MACADAM a user-friendly MetAboliC pAthway DAtabase for complex Microbial community function analysis

Malo Le Boulch, Patrice Dehais, Sylvie Combes, Géraldine Pascal

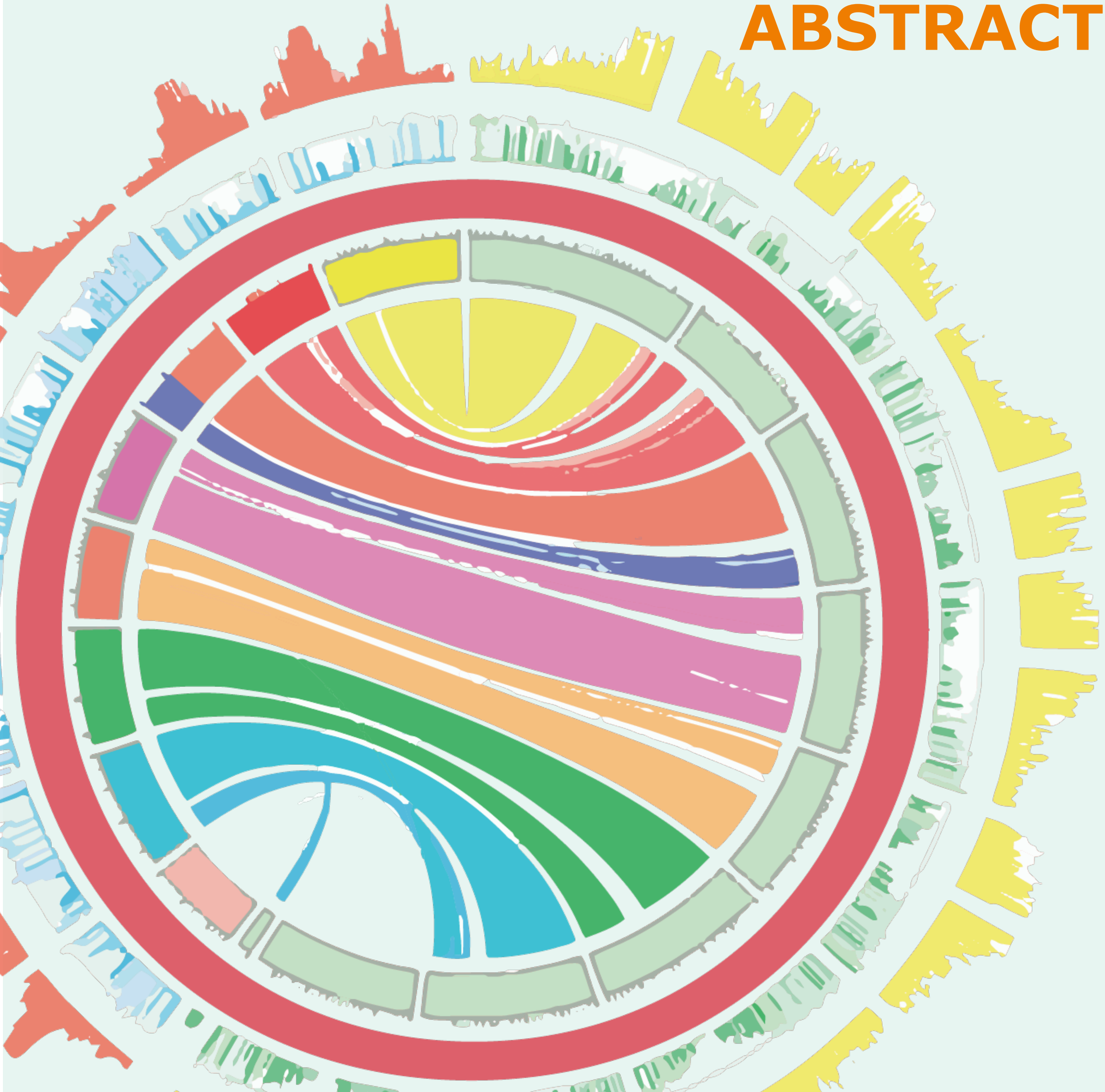
JOURNÉES OUVERTES DE BIOLOGIE INFORMATIQUE & MATHÉMATIQUES

JOBIM 2018

03>06 JUIL

PALAIS DU PHARO | MARSEILLE

ABSTRACTS



Marseille, June 20, 2018

Welcome to the 480 attendees of the 2018 JOBIM edition!

Eleven years after the first JOBIM edition organized in Marseille, here we are again. Since then, the bioinformatics field and its community have notably grown up as reflected by the numbers of attendees and submissions: we have received 263 abstracts for communications that will be presented as 46 talks and 217 posters.

This year, the conference is divided in five main sessions, that will be all opened by a keynote talk by **Ludovic Orlando**, **Pierre-Antoine Gourraud, Eleftheria Zeggini, Edda Klipp**, **Emmanuel Levy** and **Elizabeth Purdom**, respectively. We tremendously thank them to have accepted to participate and to contribute to the success of this Marseilles edition.

In addition to the simplified submission format, the novelties this year lie in the diversity of the poster tracks to highlight and cover the work of the whole community (Research, Service and Platform activity, Thematic networks, Working groups and Associations) and in the return of the flash poster presentations.

We sincerely thank all the members of the Program Committee who helped us to set up such a great program and reviewed all submissions in time. This task would have been impossible without you!

We are indebted to the organizing institutions and all our partners and sponsors for their financial supports.

For their daily efforts for the past 18 months, we thank the local Inserm administration, and Christa Roqueblave Conseil and Myriam Ramadour for their unfailing professionalism and assistance...Finally, many thanks to the local organizing committee.

Enjoy JOBIM 2018!

Christine Brun & Benoit Ballester

ORGANIZED BY

sfbi
société française de bioinformatique

Inserm
La science pour la santé
From science to health

cnrs

Aix*Marseille université
Initiative d'excellence

ifb
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

GdR BiM

PARTNERS & SPONSORS

A*Midex
Initiative d'excellence Aix-Marseille

aviesan
alliance nationale pour les sciences de la vie et de la santé

cancéropôle
Provence-Alpes-Côte d'Azur
le propulseur régional des recherches et innovations anticancers

EMBL-EBI

FONDATION ARC POUR LA RECHERCHE SUR LE CANCER
Reconnue d'utilité publique

FRANCE GÉNOMIQUE

GDR GÉNOMIQUE ENVIRONNEMENTALE

GenoScreen
Innovative Genomics

GRAND LUMINY Technopôle

INRA
SCIENCE & IMPACT

INSTITUT PAOLI-CALMETTES
Centre régional de lutte contre le cancer Provence-Alpes-Côte d'Azur

OMIC TOOLS

RÉGION SUD
PROVENCE ALPES CÔTE D'AZUR

DÉPARTEMENT BOUCHES-DU-RHÔNE

VILLE DE MARSEILLE
www.marseille.fr

Sponsors posters

Facilitating bioinformatics analyses by combining artificial and collective intelligence

Recent breakthroughs in biological technics have generated massive amounts of complex data in biology and health. Handling, processing and storing information have become new challenges for biologists. Addressing the challenge of identifying the best resources to handle biological big data, we created a decision-making web platform for bioinformatics resources. To do so, we followed a three-fold approach:
First of all, we built a database for bioinformatics resources fed by manual curation. Its specificity lies in a categorization of resources by application and analysis step. Each tool possesses its own page featuring technical and scientific information as well as community feedback.
Then, we added a layer based on artificial intelligence to automatically display optimal pipelines of tools for any biological question asked via the integrated search engine. They are generated using a combination of machine learning, manual curation and community feedback. The sequence can evolve in real-time if suggested tools are switched or optional analysis steps added. We also implemented various filters to match the user IT skills level or favorite operating system for example.
Ultimately, we developed a feature to easily run analysis on a secured cloud from the generated pipelines. When uploaded, files are analyzed beforehand to ensure compatibility, then required software are installed on containers before being used for the analysis.
By bridging artificial and collective intelligence, we offer a solution to facilitate bioinformatics analyses in a rapidly evolving field.

https://omictools.com/

GenoScreen is a French biotech company that specializes in genomics and bioinformatics.

Since 2001, we have offered innovative services and solutions (Based on the characterization and exploitation of DNA/RNA) to research groups from the private and public sectors.
GenoScreen works on all kinds of genomes (Human, animal, plant and microbial genomes) and delivers solutions for human healthcare, the agrifood industry, the environmental sector, and the cosmetics industry.
Our R&D and Innovation department's highly qualified staff are committed in the development of dedicated molecular and analytical solutions in the field of microbial genomics. GenoScreen constantly interacts with top-level academic research teams, in order to maintain its cutting-edge expertise and offer high-quality services and products to its customers.

https://www.genoscreen.fr/

Table of contents

Bioinformatics for Bugs, Beasts and Greens

Posters

New challenges for bioinformatics in the personalized medicine era 121

Oral Communications

Posters

Posters

Oral Communications

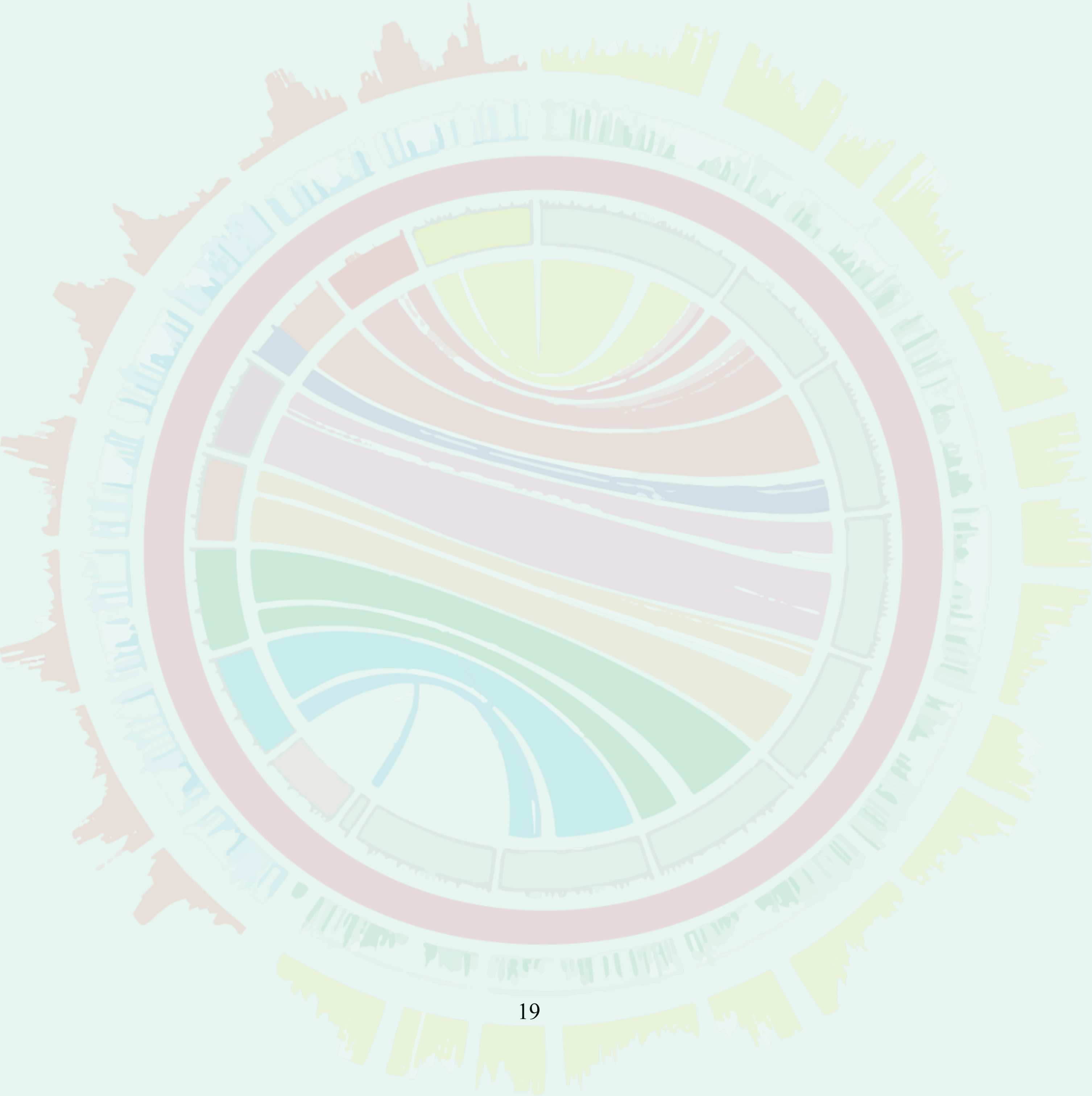
Posters

Data Science **505**

Oral Communications

Posters

Bioinformatics for bugs, beasts and greens



The evolution of large and giant viruses and their relationships with Eukaryotes

Julien Guglielmini * [1], Patrick Forterre† [2,3], Morgan Gaia‡ [4]

[1] Hub Bioinformatique et Biostatistique - Bioinformatics and Biostatistics HUB – Institut Pasteur [Paris], Centre National de la Recherche Scientifique : USR3756 – C3BI, 25-28 rue du Docteur Roux, 75724 Paris cedex 15, France

[2] Institut de Biologie Intégrative de la Cellule (I2BC) – Université Paris-Sud - Paris 11, Commissariat à l'énergie atomique et aux énergies alternatives : DRF/I2BC, Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR9198 – Bâtiment 21, 1 avenue de la Terrasse, 91198 Gif/Yvette cedex, France

[3] Biologie Moléculaire du Gène chez les Extrêmophiles (BMGE) – Institut Pasteur [Paris] – Département Microbiologie - 25-28 rue du Docteur Roux, F-75724 Paris Cedex 15, France

[4] Biologie Moléculaire du Gène chez les Extrêmophiles (BMGE) – Institut Pasteur de Paris – 28 Rue du Docteur Roux F-75724 Paris Cedex 15, France

Since their discovery, giant viruses have regularly drawn attention from the scientific community on their very unusual features for the viral world. They all belong to the NucleoCytoplasmic Large DNA Viruses (NCLDV) group, which also encompasses large eukaryotic DNA viruses from various families associated to a variety of pathological, agricultural, and environmental concerns, such as Variola (*Poxviridae*), swine fever (*Asfarviridae*), and fish-killing algal bloom termination (*Phycodnaviridae*).
The evolution of this group is not clearly understood, with the few studies on this topic offering very contrasted results and controversies, notably over their potential relationships with Eukaryotes. NCLDVs do indeed share proteins with other organisms including Eukaryotes. The origin of these proteins is not clear: were they transferred from pre-existing viruses to cellular organisms, or were they stolen by viruses from their hosts, as regularly framed by the virus pickpocket hypothesis? Some even suggested that they could be relics from a 4th domain of life that shrank to giant parasitic viruses.

NCLDVs are large, but only some of them are considered giant, which means that their viral particle is at least 300/400nm long, and their genome at least 500/600kb (those criteria are usually acknowledged yet not officially). So far, the only group that is exclusively composed of giant viruses is the *Mimiviridae*. Other giant viruses have been hypothesized to be related to *Phycodnaviridae* while other remain unclassified. All in all, the origin of "giantism" is yet to be understood.

To get more insights into the evolution of NCLDVs, we performed comparative genomics analyses on a set of almost 100 NCLDV genomes. Using classical Blast Bidirectional Best Hit (BBH) approach, we obtained a set of proteins conserved among NCLDVs (core proteins) and

[*] Speaker

[†] Corresponding author: patrick.forterre@pasteur.fr

[‡] Corresponding author: morgan.gaia@pasteur.fr

compared them with previous published analyses. Using these results, we performed in-depth phylogenetic analyses of selected core proteins and notably of DNA-dependent RNA polymerase (RNAP) large subunits.

We could show that the concatenation of different combinations of core proteins yielded congruent results, suggesting that the core proteins have undergone a congruent evolutionary history. This first conclusion discards the hypothesis that NCLDVs are mere pickpockets stealing all their gene content from their hosts and being by-products of cellular evolution. Instead, the core genes studied here were inherited by all extant NCLDV families from a common ancestor.

By concatenating the 8 most conserved core proteins, we obtained a tree which is likely to represent a species tree for NCLDVs. The robustness of this tree is guaranteed by different state of the art methods for such deep analysis that we used (Bayesian inference under CAT-GTR model, Maximum Likelihood with mixture models). This new NCLDV tree, the most robust to date, offers a new vision of the deepest bifurcations in these viruses' evolution, allowing to delimit families and to update their classification.

Importantly, our RNAP trees suggest that NCLDVs' diversification predated that of modern eukaryotes and show that RNAP evolutionary history could be more complex than previously thought, potentially evolving transfers from virus to cells and/or vice-versa.

Keywords: Giant viruses, Evolution, Phylogeny

Co-option of complex molecular system in bacterial and archaeal membrane

Rémi Denise [*][†] [1], Sophie Abby [2], Eduardo Rocha [3]

[1] Génomique évolutive des Microbes – Institut Pasteur [Paris], CNRS : UMR3525 – Département Génomes et Génétique - 25-28 rue du docteur Roux, F-75724 Paris Cedex 15, France
[2] Techniques de lÍngénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications [Grenoble] (TIMC-IMAG) – Centre National de la Recherche Scientifique : UMR5525, Université Grenoble Alpes – Domaine de la Merci - 38706 La Tronche, France
[3] Génomique évolutive des Microbes – Institut Pasteur [Paris], CNRS : UMR3525 – Département Génomes et Génétique - 25-28 rue du docteur Roux, F-75724 Paris Cedex 15, France

Protein secretion systems, and related macromolecular complexes, are present in many bacterial and archaeal species where they are involved in virulence, scavenging, and other biotic and abiotic interactions. These systems are made of many different proteins that interact together to allow other proteins to pass through the cell envelope (and in some cases into other cells). These systems are very interesting from the evolutionary point of view. First, their proteins have very diverse evolutionary rates and patterns of conservation, in spite of being part of the same machinery. Second, their genetic organization reflects structural interactions. Third, many of these systems were co-opted in complex evolutionary processes from other molecular structures involved in other types of functions. These processes of co-option (or exaptation) of functional or structural traits are thought to have been a major driver of functional innovation throughout the history of life.

Biochemical, phylogenetic, and structural evidence show that a family of molecular machineries including the type II secretion system (T2SS, involved in protein secretion), type IV pilus (T4P, involved in cell motility by a mechanism of pilus extension and retraction, adherence and virulence), Tad pilus (adherence and virulence), the competence apparatus (Com, involved in natural transformation) and the related pilus in Archaea (T4PArchaea, motility through a rotary motor and other unknown functions) share key homologous genes. These key components are a set of ATPases, the inner membrane platform, the major pilin, the prepilin peptidase, and a secretin. They form a small group of families that we could reconstruct using profile-profile alignments.

We designed custom comparative genomics tools to detect and distinguish these macromolecular systems in genome sequences. Components were searched using HMM protein profiles. Systems were identified and distinguished based on the quorum and organisation of these components in the genome according to pre-defined models. In addition to a model for each of the systems, we created a simplified model to detect putative novel systems related to the classical ones containing the same key homologous components.

We have detected more than 6600 systems in Prokaryotes on a dataset of 5776 complete genomes. For each key component, we inferred a phylogeny, which we reconciled with those of other compo-

[*] Speaker

[†] Corresponding author: remi.denise@pasteur.fr

nents after having taken into account the presence of paralogs. With the reconciled phylogenies, we inferred and rooted the tree of the systems. This rooted phylogeny showed that systems form monophyletic groups for each type of system. Hence, even if the systems are often gained and lost by horizontal transfer and have homologs in the other types of systems, each type evolves independently. Importantly, the rooted phylogeny allows to orient the events of co-option of machineries into other functions. For example, it clearly shows that T2SS were co-opted from T4P. This fits our observation that T2SS have narrower taxonomic distributions than T4P.

The rooted tree showed a close relationship between the Tad and the T4PArchaea, suggesting that these systems either diversified before the last common ancestor or were initially present in the Bacteria and then were transferred horizontally to the Archaea. The first hypothesis is sustained by the observation that both major branches of the systems' phylogeny include types of systems (T4P, Tad) present in all major clades of Bacteria (where usually the monoderms cluster apart from diderms). The second hypothesis is sustained by the observation of frequent transfer of these systems within phyla.

Some of the systems (notably T4P) tend to be encoded in different loci, whereas most others are usually encoded in one single locus. We put forward the hypothesis that horizontal transfer was more frequent in systems encoded in one single locus. We used amalgamated likelihood estimation (ALE) to estimate the rates of deletion, transfers and loss (DTL) and to test this hypothesis. The results suggest than the genetic organisation of the systems increase the transfer rate with more transfers found for systems encoded in single locus. We can also see than the competence systems of monoderm seems to be less transferred than the other, it may be due to the fact than close to all the monoderms have a competence system and only one, so maybe there is a selection pressure to have only one of this system that reduce the transfers rate.

Finally, our evolutionary scenario shows that competence systems of diderms are scattered in the tree of T4P, with competence apparatus of monoderms being a sister clade of the group formed by T4P and T2SS. This information and the fact than some known T4P are used in bacteria for natural transformation, shows that transformation is a very ancient mechanism. It also opens the unexpected possibility that T4P were initially dedicated to natural transformation.

The integration of genomics and phylogenetic analyses facilitated the discrimination between related systems and was then used to perfect the models used to detect the systems. This significantly improved our ability to identify (we initially detected 1584 generic systems and we could later re-class 88% of them) and discriminate these partly homologous systems in the genome data. Hence, this work exemplifies the interest of integrating comparative genomics and phylogenetics in genome annotation.

Keywords: Phylogenomic, Microbiology, Genomic, Evolution, Secretion system

A new rapid, flexible and intuitive software to simulate phylogenies of infections

Gonché Danesh * [1], Emma Saulnier [2], Samuel Alizon [1]

[1] Laboratoire "Maladies Infectieuses et Vecteurs, Ecologie, Evolution et Contrôle" (MIVEGEC) – UMR CNRS, IRD, UM – 911 Avenue Agropolis BP 64501 34394 Montpellier cedex 5 - FRANCE, France
[2] Institut de Biologie Computationnelle (IBC) – CNRS : UMR5506, Université Montpellier 2 (FRANCE) – 95 rue de la Galéra, 34095 Montpellier, France

A new rapid, flexible and intuitive software to simulate phylogenies of infections
The field of phylodynamics hypotheses that the way pathogens spread leaves footprints in their genomes [1]. Phylodynamic methods have made great progress in the past decade, especially through the popular software package BEAST [2], but the increase of the size of phylogenies and the complexity of models has led to the development of inference approaches that require simulating transmission trees [3][4]. Although these simulators are essential for phylodynamic inference, they are rarely compared.

Here we introduce a new simulator developed in Rcpp composed of two modules : one for simulating the trajectory of the epidemic according to a specific compartmental model, based on Gillespie's Stochastic Simulation Algorithms [5] and two of its variants, and one for simulating sampled transmission tree from an epidemiological trajectory using the coalescent approach. We compared the performances of our simulator, the rcolgem R package which is based on the Adam's approximation [3][6] and the software MASTER which uses Gillespie's algorithms [7].

By simulating epidemics using either a simple Susceptible-Infected-Recovered (SIR) model or a more detailed model describing HIV spread in a heterogeneous population, we find that the nature of the optimal method in terms of computation time depends on the number of phylogenies simulated, the parameter values and the complexity of the model. We also show that our method usually outperforms existing ones, in terms of rapidity and sometimes even in terms of accuracy (using MASTER as a reference for the trajectory and the phylogeny), when simulating a large number of trees.

References:

Grenfell BT, Pybus OG, Gog JR, *et al.* (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science, 303(5656): 327-332

Drummond AJ, Rambaut A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214

Volz EM (2012) Complex population dynamics and the coalescent under neutrality. Genetics.

*Speaker

190(1):187-201

Saulnier E, Gascuel O, Alizon S. (2017) Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. PLoS Computational Biology. 13(3):e1005416

Keeling MJ, Rohani P. (2008) Modeling infectious diseases in humans and animals. Princeton.

Rasmussen DA, Volz EM, Koelle K. (2014) Phylodynamic inference for structured epidemiological models. PloS Comput Biol. 10(4):e1003570

Vaughan TG, Drummond AJ. (2013) A stochastic simulator of birth-death master equations with application to phylodynamics. Mol Biol Evol. 30(6):1480–1493

Keywords: Phylodynamics, Virus, Infectious diseases, Epidemiology, Phylogenetics

Combining eDNA metabarcoding and supervised machine learning for routine environmental applications: an example with marine aquaculture

Tristan Cordier * [1]

[1] Université de Genève (UNIGE) – Sciences III. Bd d'Yvoy 4, 1205 Genève, Switzerland

Monitoring natural resources and ecosystem services is of prime importance for the sustainable development of human societies. Environmental impacts of human activities in marine ecosystems, such as aquaculture, oil drilling or mining operations, are traditionally assessed through benthic biodiversity surveys. This involves the sorting and the taxonomic identification of thousands of macroinvertebrates specimens, which is time consuming and taxonomic-expertise demanding. The development of high-throughput DNA sequencing has paved the way toward fast and objective description of biological communities. However, usually more than half of eDNA sequences remain unassigned or belong to taxa of unknown ecology, which prevent their use for inferring the ecological quality status of a sample. We recently showed that supervised machine learning (SML) proved efficient for the building of robust predictive models from eDNA metabarcoding data, regardless of the taxonomic assignments of the sequences. This allows to use almost all the sequences of a dataset, constituting a more holistic approach to infer the ecological quality status. Combining eDNA high-throughput sequencing and SML holds the potential to overcome the limitations of macroinvertebrates inventories. **We will present recent results and future directions toward combining eDNA metabarcoding data and machine learning for the routine biomonitoring of marine environments.**

Keywords: metabarcoding, machine learning, biomonitoring

*Speaker

MACADAM a user-friendly MetAboliC pAthway DAtabase for complex Microbial community function analysis

Malo Le Boulch *† [1], Patrice Dehais [1], Sylvie Combes [1], Géraldine Pascal [1]

[1] GenPhySE - UMR 1388 (Génétique Physiologie et Systèmes dÉlevage) – Université de Toulouse, INRA, INPT, ENVT – 24, chemin de Borde-Rouge -Auzeville Tolosane31326 Castanet Tolosan, France

The progress of genome sequencing and in bioinformatics has opened new possibilities in particular to link genome annotation to functional information like metabolic pathways. This link is based on the development of functional databases that can link genome annotations to functional annotations. Nowadays, there are multiple generalist databases like KEGG(1), MetaCyc(2), HMDB(3), WikiPathways(4) or Reactome(5). However, most of them focus only on few highly curated species of interest whereas the others have moved to a subscription system or cannot be requested on a completely taxonomic rank, particular function or even a compound constituting a metabolic pathway.
Here we propose MACADAM (MetAbolic pAthway Database for complex Microbial communities) a user-friendly functional database. This database is based on high quality genomes from RefSeq(6) (https://www.ncbi.nlm.nih.gov/refseq/), MicroCyc(7) (http://www.genoscope.cns.fr/agc/microscop Faprotax(8) (http://www.zoology.ubc.ca/louca/FAPROTAX/lib/php/index.php) and the IJSEM phenotypic database (9) (https://doi.org/10.6084/m9.figshare.4272392.v3). To ensure high quality data, we kept from RefSeq only genomes with no sequencing errors. In order to discover the functional potential of each one of these genomes we used the Pathway-Tools(10) software from the Metacyc metabolic pathway database. Pathway-Tools links the functional annotation of the genomes of interest to metabolic pathways in the Metacyc database and produces a PGDB file as an output (Pathway-Genome DataBase). A PGDB, which includes all metabolic pathways found via the genome annotation, is computed for each organism. To increase the exhaustiveness and the quality of the functional annotation of the genomes in the MACADAM database, we completed our database with the MicroCyc database. MicroCyc gathers a collection of PGDBs created by the LABGeM Genoscope The MicroCyc database is highly curated by biologists and by bioinformatics methods which increases the quality of the PGDBs. Finally, we also included the FAPROTAX (Functional Annotation of Prokaryotic Taxa)(8) and IJSEM phenotypic database(9) databases which are manually curated functional databases. To link the PGDBs to a taxonomic lineage, we used the NCBI taxonomy which is the most interoperable taxonomic database(11). Pathway hierarchies, compounds, enzymatic reactions and enzymes were in the database.

For each pathway of each PGDB, we computed a Pathway Score (PS) that reflects the functional potential of a given organism to trigger a defined metabolic pathway. If the score is near

[*] Speaker

[†] Corresponding author: malo.leboulch@inra.fr

0, few enzymes catalysing reactions of the pathway of interest are annotated in the organism's genome. If this score is equal to 1, all the pathway enzymes are detected and consequently the pathway is thus complete. Some constituting enzymes can be present several times in the organism's genome. To take into account this latter information we calculated the Frequency Pathway Score that corresponds to the number of enzymes present in the genome for a defined pathway divided by the total number of enzymes needed to achieve the pathway.

To optimize and facilitate the request of the database we used SQLite (https://www.sqlite.org), a relational database management system. MACADAM accepts requests with a taxonomic name optionally associated with a taxonomic rank. MACADAM also provides the possibility to query directly on a metabolic pathway name, a compound or an enzymatic reaction. All the field search texts can be filled with complete or incomplete strings. The database query can be done via one or more complete or incomplete taxonomic words and, optionally, a specific taxonomic rank. MACADAM database accepts requests on several bacterial clade at the same time in order to explore functional diversity among several clade. Additionally, a PS value can be specified so that, for example, only complete pathways are queried (PS=1). Considering the impact of database obsolescence on biological analysis(12), the MACADAM construction script can be easily launched in order to keep data up to date.

The MACADAM database contains 9146 different organisms including 3805 organisms from the FAPROTAX database. It contains 1198 unique metabolic pathways as well as 82 functional annotation by FAPROTAX. Additionally, MACADAM contains 2313 compound names, 406 reactions names and 7620 enzyme names.

MACADAM includes all phyla recognized by the LSPN (List of Prokaryotic names with standing in nomenclature) (http://www.bacterio.net/) and the 10 other newly proposed phyla from the NCBI taxonomy. Proteobacteria is the most prevalent phylum and accounts for more than 50% of the collected genome included in MACADAM followed by the Firmicutes phylum with more than 20% of genomes. We explain this pre-eminence by the research effort devoted on these phyla by biologists. In agreement, Bacillus and Escherichia represent the most genera in the database (7% and 6.3% of the database respectively). The most prevalent functions in MACADAM belong to the Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis class (15%), Amino Acids Biosynthesis class (9,7%) and Amino Acids Degradation (7,3%).

The output is in the form of a tsv file. The matching point in the taxonomy of the inputed word is displayed as well as the corresponding pathways. If there is no functional information available for a given taxonomic name then functional information of the upper taxonomic rank is provided.

The output file also contains the number of times a pathway is present in the overall organism selected. This allows finding out if a specific metabolic pathway is shared in a bacterial clade or if the pathway is present in few organisms or specific strains. The metabolic hierarchy is also provided. For an easy and intuitive consultation of MACADAM database, without using command lines, a website is currently under construction. The creation of this database is a first step in the exploration process of bacteria. Later, a tool will use it to infer functional potential of bacterial communities.

MACADAM makes it possible to obtain any bacteria functional potential from a taxonomic name. MACADAM makes sense in the context of metabarcoding data where the key OTUs have been identified in a bacterial community. MACADAM helps to retrieve the functional potential of these OTUs.

References:

1. Kanehisa, M., Furumichi, M., Tanabe, M., et al. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Res., 45, D353–D361.

2. Caspi, R., Billington, R., Fulcher, C. A., et al. (2018) The MetaCyc database of metabolic pathways and enzymes, Nucleic Acids Res., 46, D633–D639.

3. Wishart, D. S., Feunang, Y. D., Marcu, A., et al. (2018) HMDB 4.0: the human metabolome database for 2018, Nucleic Acids Res., 46, D608–D617.

4. Slenter, D. N., Kutmon, M., Hanspers, K., et al. (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research, Nucleic Acids Res., 46, D661–D667.

5. Fabregat, A., Jupe, S., Matthews, L., et al. (2018) The Reactome Pathway Knowledgebase, Nucleic Acids Res., 46, D649–D655.

6. O'Leary, N. A., Wright, M. W., Brister, J. R., et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, Nucleic Acids Res., 44, D733-745.

7. Vallenet, D., Calteau, A., Cruveiller, S., et al. (2017) MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes, Nucleic Acids Res., 45, D517–D528.

8. Louca, S., Jacques, S. M. S., Pires, A. P. F., et al. (2017) Functional structure of the bromeliad tank microbiome is strongly shaped by local geochemical conditions, Environ. Microbiol., 19, 3132–3151.

9. Barberán, A., Velazquez, H. C., Jones, S., et al. (2017) Hiding in Plain Sight: Mining Bacterial Species Records for Phenotypic Trait Information, *mSphere*, **2**, e00237-17.

10. Karp, P. D., Latendresse, M., Paley, S. M., et al. (2016) Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology, Brief. Bioinform., 17, 877–890.

11. Balvočiūtė, M. and Huson, D. H. (2017) SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare?, BMC Genomics, 18, 114.

12. Wadi, L., Meyer, M., Weiser, J., et al. (2016) Impact of outdated gene annotations on pathway enrichment analysis, Nat. Methods, 13, 705–706.

Keywords: Pathway Database, Metabolic Pathway

Probabilistic PCA for count data in microbial ecology

Mahendra Mariadassou *† [1], Stéphane Robin [2], Julien Chiquet [3]

[1] Institut National de Recherche Agronomique - Centre de Jouy-en-Josas (Unité MaIAGE) – Institut national de la recherche agronomique (INRA) : UR1404, Université Paris-Sud - Université Paris-Saclay – Domaine de Vilvert 78352 JOUY-EN-JOSAS CEDEX, France

[2] UMR MIA-Paris – AgroParisTech, Institut National de la Recherche Agronomique - INRA, Université Paris-Saclay – 16 rue Claude Bernard F - 75 231 Paris Cedex 05, France

[3] AgroParisTech (AgroParisTech) – Institut national de la recherche agronomique (INRA) : UMR518, AgroParisTech – AgroParisTech 75231 Cedex 05, Paris - France, France

Many application domains such as ecology or genomics have to deal with multivariate non Gaussian observations. A typical example is the joint observation of the respective abundances of a set of species in a series of sites, aimed at understanding the co-variations between these species. Metabarcoding experiments resort to this situation and, as the number of observed species is large, a first task is often to summarize the information in a lower dimension for vizualization or primary inspection. Principal Component Analysis (PCA) is a popular tool to make this reduction; it relies on a low-rank approximation of the covariance matrix. The Gaussian setting provides a canonical way to model the dependencies between species, but does not apply in general to such data.

We consider here the multivariate exponential family framework for which we introduce a generic hierarchical model with multivariate Gaussian latent variables. Briefly, and with the specific application of metabarcoding in mind, we use a hierarchical Poisson log-normal (PLN) model with a latent Gaussian layer and an observed Poisson layer. We use the multivariate Gaussian formulation to model a simple low-rank representation of the samples in the latent space and then consider conditionally independent univariate Poisson counts in the observation layer.

This model can be seen as an extension of probabilistic Principal Component Analysis (pPCA) to non

[*] Speaker

[†] Corresponding author: Mahendra.Mariadassou@jouy.inra.fr

Gaussian settings and enables us to account for covariates and offsets at little additional cost. This is particularly important in metabarcoding to correct for different sequencing depths and abiotic conditions in all samples. We introduce covariates in the Poisson layer, using the Generalized Linear Model (GLM) framework, to control for confounding factors (pH, salinity, etc) and more generally all auxiliary variables recorded on the samples.

Unlike the purely Gaussian setting, the likelihood is generally not tractable in this framework. We resort instead to a variational approximation to obtain a tractable lower bound of the likelihood. We then use gradient descent to solve the corresponding optimization problem. Finally, we develop a model selection criterion to select the number of dimension of the latent space.

Formal expression of the gradient depends on the exponential family at hand and does not always have a analytic expression. It can however always be evaluated efficiently using Gauss-Hermite quadrature of one dimensional integrals. Moreover, in the specific case of Poisson-log-normal models, both the lower bound of the likelihood and its gradient have close form expressions and the inference procedure is efficient.

We then turn our attention to two published metabarcoding datasets: one the impact of weaning on pig gut microbiota (Mach et al. 2015) and the other on the impact of a fungal pathogen on the phyllosphere of oaks (Jakuschkin et al. 2016). We show that the PLN framework is flexible enough to simultaneously analyze metabarcoding data obtained on the same samples but with different markers (16S for the bacterial fraction, ITS for the fungal fraction, etc). We also show that PLN PCA (i) scales nicely to hundreds of samples with hundreds of OTUs, (ii) recovers well documented structures and (iii) is able to discover subtle second-order structuring effects that are masked by stronger first order effects, thanks to the introduction of covariates.

Poisson log-normal PCA (PLN-PCA) is presented with full mathematical details in (Chiquet, Mariadassou, and Robin to appear) and implemented as an R package available from https://github.com/jchiquet/PLNmodels

- Chiquet, J., M. Mariadassous, and S. Robin. n.d. "Variational Inference for Probabilistic Poisson Pca." arXiv Preprint, to Appear in Annals of Applied Statistics. https://arxiv.org/abs/1703.06633.
- Jakuschkin, B., V. Fievet, L. Schwaller, T. Fort, C. Robin, and C. Vacher. 2016. "Deciphering the Pathobiome: Intra-and Interkingdom Interactions Involving the Pathogen Erysiphe Alphitoides." Microbial Ecology. Springer, 1–11.
- Mach, N., M. Berri, J. Estellé, F. Levenez, G. Lemonnier, C. Denis, J.-J. Leplat, et al. 2015. "Early-Life Establishment of the Swine Gut Microbiome and Impact on Host Phenotypes." Environmental Microbiology Reports 7 (3). Wiley-Blackwell: 554–69. doi:10.1111/1758-2229.12285.

Keywords: metabarcoding, dimension reduction, count data, visualization

Single-cell genome study of marine protists in the framework of the Tara Oceans project

Léo D'agata [*] [†] [1], Yoann Seeleuthner[‡] [2], Artem Kourlaiev[§] [1], Marc Wessner[¶] [1], Benjamin Noel[‖] [1], Olivier Jaillon[**] [2], Julie Poulain[††] [1], Patrick Wincker[‡‡] [2], Jean-Marc Aury [1]

[1] Commissariat à l'Energie Atomique (CEA), Genoscope, Institut de Biologie François-Jacob – CEA Evry 2 rue Gaston Crémieux 91006 Evry cedex – F-91057 Evry, France, France

[2] Génomique métabolique (UMR 8030) – Commissariat à l'énergie atomique et aux énergies alternatives : DRF/IG, Université d'Évry-Val-d'Essonne, Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR8030 – GENOSCOPE, 2 rue Gaston Crémieux 91057 Evry Cedex, France

Marine plankton is responsible for more than half of the oxygen production on the Earth. Nevertheless, it remains little known and the impact of global warming and pollution on him is still unclear. Its contribution in CO2 absorption and climate regulation could indeed be directly affected. In order to answer these questions, the *Tara* expedition, in which Genoscope participates, took place between 2009 and 2013 under the name *Tara* Oceans [1].

One of the aims is the precise genome study of the marine protists, and their geographical distribution. metagenomic and metatranscriptomic samples were collected during the expedition. But since these protists were mostly not suitable for laboratory cultivation, it was also necessary to use single-cell sequencing. The corresponding protocol consisted of extracting the DNA from a single cell, amplifying it by MDA (Multiple Displacement Amplification) [2] and finally sequenced it. The extraction and amplification steps, however, generate many problems that make assembling their genome difficult. Some regions may indeed be uncaught during DNA extraction while others are not amplified. The amplification applied to these protists also generates a very irregular coverage. Appropriate tools are therefore needed to assemble these genomes.

A specific assembly workflow has been set up at Genoscope to process this data. One idea proposed by this workflow is to use the synergy of several cells to solve as much as possible the problem of uncovered regions. However this solution known as "co-assembly" requires knowing in advance which cells belong to the same organism. 18S ribosomal DNA (marker gene in eukaryotes) may provide an early response. However, this sequence is sometimes not very specific and does not fully reflect the genomes diversity. To resolve this problem, we used the method

[*] Speaker

[†] Corresponding author: leo.dagata@gmail.com

[‡] Corresponding author: yseeleuthner@genoscope.cns.fr

[§] Corresponding author: akourlai@genoscope.cns.fr

[¶] Corresponding author: mwessner@genoscope.cns.fr

[‖] Corresponding author: bnoel@genoscope.cns.fr

[**] Corresponding author: ojaillon@genoscope.cns.fr

[††] Corresponding author: poulain@genoscope.cns.fr

[‡‡] Corresponding author: pwincker@genoscope.cns.fr

Corresponding author: jmaury@genoscope.cns.fr

ANI (Average Nucleotide Identity) [3] which makes it possible to evaluate the similarity between several genomes by making a pairwise comparison of their assembly. The identity rate obtained between the cells informs us about their proximity or distance. About one hundred assemblies of single-cell amplified genome (SAG) were obtained, from which one-third were co-assemblies.

A specific structural annotation workflow was then applied to the resulting assemblies. This workflow consists of constructing gene models from three different sources. The first source comes from the protein alignment result. The second source comes from the result of an Ab-initio gene prediction tool (trained using gene models produced with protein alignments). Next reads from metatranscriptomic samples of the Tara Oceans expedition were aligned and used as input to predict genes using the Gmorse software [4]. Finally, the three data sources were combined using the Gmove tool [5] to produce the final gene catalogue of each SAG.

Then, an analysis of the biogeographical distribution was also conducted. This analysis consists in determining the abundance of each SAG according to the origin of the metagenomic samples. The conduct of this method is first to align the metagenomic data on the assemblies and then to calculate their abundance by dividing the total number of reads aligned on the genome by the total number of reads present in the sample. This abundance, however, did not take into account the completeness of the assembly obtained with respect to the real genome size. Gene completion was therefore calculated for each assembly, and used to extrapolate previous abundance results to estimate a more accurate biogeography.

References:

Eric Karsenti, Silvia G Acinas, et al. (2011) A holistic approach to marine eco-systems biology.

Frank B Dean, Seiyu Hosono, et al. (2002) Comprehensive human genome amplification using multiple displacement amplification.

Mangot, J.-F., R. Logares, et al. (2017) Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells.

Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F. Annotating genomes with massive-scale RNA sequencing. Genome Biol. 2008;9(12):R175. doi: 10.1186/gb-2008-9-12-r175. Epub 2008 Dec 16. PubMed PMID: 19087247; PubMed Central PMCID: PMC2646279.

Dubarry, M., Noel, B., & Rukwavu, T., & Aury, J. M. (2016) Gmove a tool for eukaryotic gene predictions using various evidences.

Keywords: Single cell, Sequencing, Coassembly, Structural annotation, Marine protists, Tara Oceans, Metagenomic, Metatranscriptomic, Phylogeny, Biogeography

PPanGGOLiN: Depicting microbial diversity via a Partitioned Pangenome Graph

Guillaume Gautreau *† [1], Christophe Ambroise [2], Catherine Matias [3], Amandine Perrin [4], Valentin Sabatet [1], Rémi Planel *

[1], Marie Touchon [4], Claudine Médigue [1], Eduardo Rocha [4], Stéphane Cruveiller [1], David Vallenet‡ [1]

[1] Laboratoire d'Analyse Bioinformatique en Génomique et Métabolisme (LABGeM) – CEA, Genoscope : DRF/IBFB/Gen, CNRS : UMR8030, Université d'Evry - Université Paris-Saclay – 2 rue Gaston Crémieux, France
[2] Laboratoire de Mathématiques et Modélisation d'Evry – CNRS : UMR8071, Université d'Evry-Val d'Essonne, Institut national de la recherche agronomique (INRA) – France
[3] Laboratoire de Probabilités et Modèles Aléatoires (LPMA) – Université Pierre et Marie Curie (UPMC) - Paris VI, CNRS : UMR7599, Université Paris VII - Paris Diderot – France
[4] Génomique évolutive des Microbes – Institut Pasteur [Paris], Centre National de la Recherche Scientifique : UMR3525 – Département Génomes et Génétique - 25-28 rue du docteur Roux, F-75724 Paris Cedex 15, France

Motivations:

By collecting and comparing all the genomic sequences of a species, pangenomics studies focus on overall genomic content to understand genome evolution both in terms of core and accessory parts (Tettelin et al. 2005). The core genome is defined as the set of genes shared by all the organisms of a taxonomic unit (generally a species) whereas the accessory part (also named, variable regions or peripheral regions) is crucial to understand the adaptive potential of bacteria and contains genomic regions that can be exchanged between strains by horizontal transfers (i.e. the mobilome, Frost et al. 2005).

Core genes are most often defined as the set of ubiquitous genes in a clade (Tettelin et al. 2005 and Vieira et al. 2011). However, this definition has 2 major flaws : (i) it is not robust against poorly sampled data because it is highly reliant on the presence or absence of a single organism; (ii) it misses many core genes (false negatives) because of the high probability to lose at least one of the core genes due to sequencing, assembling or annotation artifacts. Potential presence in the dataset of variants missing a gene because the associated function is socialized [sic] in a community (see the Black Queen Hypothesis, Morris et al. 2012) can also drop down the core genome. As pointed out by (AcevedoRocha et al. 2013), "functional ubiquity cannot be equated to sequence/structural ubiquity", the core genome definition has thus been pointed

*Speaker

†Corresponding author: guillaume.gautreau@free.fr

‡Corresponding author: vallenet@genoscope.cns.fr

out as too conservative for being useful (Tonder et al. 2014). As a consequence, (AcevedoRocha et al. 2013) propose to rather focus on "persistent" genes, namely genes that are conserved in a majority of genomes. Some equivalent words to 'persistent' have also been introduced as 'soft core' (ContrerasMoreira et al. 2013) or 'extended core' (Lapierre et al. 2009, Bolotin et al. 2017), 'stabilome' (Vesth et al. 2010). This definition advocates for the use of a threshold on the frequency of appearance of a gene among the species of a clade, above which the gene is declared as a persistent one (generally gene families present at least in a range comprised between 90% and 95%). This approach gives an attractive answer to the issues raised by the original definition of the core genome but nevertheless has its own disadvantage that lies in choosing an appropriate threshold.

Moreover, the usual dichotomy between core and accessory genome does not faithfully report the diverse ranges of gene frequencies in a pangenome. The gene frequency distribution in pangenomes is extensively documented (Lapierre et al. 2009, Collins et al. 2012, Lobkovsky et al. 2013, Bolotin et al. 2015, Bolotin et al. 2017). These studies argue for the existence of an equilibrium between genes acquisition and genes loss leading to an asymmetric U-shaped distribution of gene frequencies regardless of the phylogenetic level and the clade considered (with the exception of the non-homogeneous species (Moldovan et al. 2018)). The U left, bottom and right sides correspond respectively to the rare, moderately present and highly frequent gene families. Thereby, as proposed by (Koonin,2008) and formally modeled by (Collins et al. 2012), the pangenome can be split into 3 groups.This choice helps to shed light on genes putatively associated with positive environmental adaptations while avoiding to confound them with potentially randomly acquired ones. For that purpose, the partitioning approach that we propose here divides the pangenome into (1) persistent genome, equivalent to a relaxed core genome (genes conserved in almost all genomes); (2) shell genome, genes having intermediate frequencies corresponding to moderately conserved genes potentially associated to environmental adaptation capabilities; (3) cloud genome, genes found at a very low frequency. We tackle this challenge in the present work by first proposing a method to select this threshold automatically.

Beyond the partitioning approach, the technological shifts of the sequencing methods offer us thousands of genome strains available in databases for numerous bacterial species. The processing of so many genomes poses a critical computational problem because it is no longer possible to handle comparative genomics studies as in the 90's, even with modern computing facilities. For instance, studying patterns of gene gains and losses in the evolution of a lineage is a basic question in comparative genomics but this task becomes tremendously harder when thousands of genomes have to be analyzed. Nevertheless, the information encoded in these genomes is highly redundant making it possible to design new compact ways of representing and manipulating this information. As suggested (Chan et al. 2015 and Marshall et al. 2016), a consensus representation of multiple genomes would provide a better analytical framework than using individual reference genomes. This proposition leads to a paradigm shift from the usual linear representation of reference genomes to a representation as pangenome graphs bringing together all the different known variations as multiple alternative paths. Some approaches have been developed aiming at factorizing pangenomes at the sequence level (PanCake : Ernst et al. 2013, SplitMEM : Marcus et al. 2014). However, these approaches lack direct information about genes, complicating the functional analyses from the study of the graph. Here, we introduce an extension of the concept of pangenome graph, giving it a formal mathematical representation using a graph model in which nodes represent gene families and chromosomal neighborhood information, respectively. The method introduced here can be considered as the missing link between the usual pangenomics approach (set of unlinked gene families) and the pangenome graph at the sequence level. A detailed comparison of these 2 approaches has been reviewed in (Zekic et al. 2018). Coupled with our partitioning method, this representation could be a new standard to depict all the genomic combinations of bacterial species in a single figure.

Overview of the **P**artitioned **P**an**G**enome **G**raph **O**f **L**inked **N**eighbors method:

First, the genomes of the same species (or species cluster) are annotated before bringing homolog genes together into gene families via a all vs all protein alignment. From this data, the PPanGGOLiN method merges the chromosomal links between neighboring genes to build a graph of the neighborhood between gene families weighted by the number of genomes covering each edge. In parallel, the pangenome is modeled as a binary presence/absence matrix where the rows correspond to gene families and columns to the organisms (1 in case of presence of at least one gene belonging to this gene family, 0 in case of absence). The pangenome is then partitioned into the persistent, shell and cloud partitions by evaluating, through an Expectation-Maximisation algorithm, the best parameters of a Bernoulli Mixture Model (BMM) smoothed using a Markov Random Field (MRF) (Ambroise et al. 1997). For each partition, the BMM is composed of one mean vector of presence/absence (expected to be (11...11) for the persistent, (00...00) for the cloud and diversified for the shell) associated to a dispersion vector around the mean vector (low dispersion for the persistent and the cloud; high dispersion for the shell). Once the parameters are estimated, each gene family is associated to its closest partition according to its mean vector. As it is known that core gene families share conserved genomic organizations along genomes (Fang et al. 2008), the MRF imposes that two neighboring gene families are more likely to belong to the same partition. Therefore, the MFR penalizes unreliable partition attributed to the families compared to the partition of its neighbors in the graph (the weights of the edges account in the process). The algorithm iterates between BBM and MRF until the maximization of the overall likelihood. The strength of the topological smoothing is managed via a parameter called β (if $\beta = 0$, the smoothing is disabled and the partitioning only relies on the presence/absence matrix). At the end, the partitions are then overlaid on the neighborhood graph in order to obtain what we called the Partitioned Pangenome Graph. Thanks to this graphical structure and the associated statistical model, the pangenome is resilient to randomly distributed errors (e.g. an assembly gap in one genome can be offset by information from other genomes, thus maintaining the link in the graph).

Conclusion:

Due to the significant decreasing cost of recent sequencing technologies, the past recent years have seen the explosion of whole-genome sequencing projects (WGS), most notably for pathogenic bacteria. Using portable sequencer like ONT MinION, it is soon imaginable to obtain thousands of strains for each species because of the simplicity to sequence bacteria directly on the field. Therefore, the capture of all genomic variations of a species is no longer a wishful thinking. Before the emergence of the pangenomics, the emphasis has been on identifying polymorphism information to draw some sort of epidemiological map of the lineage(s) of interest. While this has resulted in the remarkably detailed information of epidemic strains, it is rapidly showing its major weakness since the analysis of the core genes actually provides very little information on the adaptive changes because most of them arise in the shell and cloud genomes. The approach presented here sheds light on these variations to focus on the gene gains and losses that are associated with these adaptive changes in a species. In the context of comparative genomics, drawing genomes on rails like a subway map may help biologist to compare genomes of interest to the overall pangenomic diversity. This graph-based approach to represent and manipulate pangenomes provides efficient bases for very large scale comparative genomics. The method is available as a standalone tool (https://github.com/ggautreau/PPanGGOLiN) and, as mentioned in (Vallenet et al. 2017), we are currently working on its integration in the MicroScope platform.

Keywords: pangenomics, comparative genomics, microbiology

Comparative metagenomics highlighted a core of metabolic capabilities in multiple serpentinizing ecosystems

Eléonore Frouin *† [1], Fabrice Armougom [1], Gaël Erauso [1]

[1] Institut méditerranéen dócéanologie (MIO) – Institut de Recherche pour le Développement :
UMR$_D$235, AixMarseilleUniversité : UM110, UniversitédeToulon :
UMR7294, CentreNationaldelaRechercheScientifique : UMR7294 –
–M.I.O.InstitutMéditerranéendÓcéanologieBâtimentMéditerranée163AvenuedeLuminy13288Marseille, France

Serpentinizing hydrothermal systems result from in-depth waters circulation and interactions with mantle rocks (peridotites) in slow-spreading oceanic ridges or terrestrial ophiolites. The geochemical process of serpentinization yields alkaline hydrothermal fluids (pH 9-11) enriched in hydrogen, methane and small organic molecules. The by-products of this reaction are assumed to feed chemosynthetic microbial communities **[1]**. In submarine serpentinizing environments, hydrothermal fluids precipitate upon mixing with seawater and form carbonate chimneys. From the interior of these hydrothermal chimneys to the interface with seawater, proton and redox gradients represent a rich source of abiotically produced energy and may constitute favorable conditions for life's origins **[2]**. The most studied serpentinizing system is the Lost City Hydrothermal Field at 800m depth, near the Mid-Atlantic Ridge. Sampling at deep oceanic seafloor is challenging, thus, other more accessible serpentinizing fields have been investigated, such as terrestrial ophiolites and shallow submarine sites. Interestingly, taxonomic similarities were detected among geographically distant serpentinizing ecosystems, such as taxa affiliated with genus *Hydrogenophaga*, or belonging to the *Clostridiales*, *Thermoanaerobacterales* or *Methanosarcinales* orders **[3, 4]**. Apart from these few phylotypes belonging to dominant lineages frequently identified in serpentinizing habitats, the microbial communities varied considerably from site to site. However, the difference of taxonomic pattern in distinct samples does not necessarily reflect a distinct functional pattern. This study thus aims to determine if selection pressures imposed by serpentinization reactions on endogenous microbial communities lead to convergence on functional strategies for living in these harsh environments.

We explored the links between serpentinization and associated microbial communities through comparative metagenomics of serpentinizing and non-serpentinizing hydrothermal systems that are comparable regarding pH, salinity or temperature. Twenty-one publicly available metagenomes from six serpentinizing systems (submarine and continental) and four other hydrothermal systems were retrieved from SRA and MG-RAST databases. In order to standardize data processing, all metagenomes were reprocessed with a same pipeline including assembly (IDBA-UD), annotation (Prodigal), taxonomic affiliation (DIAMOND against nr, MEGAN), functional affiliation (rpsblast against COG, GhostKOALA) and normalization (bedtools, edgeR). The similarity between taxonomic profiles was computed with the Jensen-Shannon distance and visualized

*Speaker

†Corresponding author: eleonore.frouin@mio.osupytheas.fr

using principal coordinates analysis, while the functional profiles were categorized using hierarchical clustering. Finally, specific capabilities of the microbial communities from serpentinizing ecosystems were identified using a Random Forest supervised classification.

Both taxonomic and functional profiles showed that the microbiomes of the two submarine serpentinite-hosted systems, the Prony and Lost-City Hydrothermal Fields were more distant than expected from previous 16S rRNA surveys. Specifically, the microbial biosphere of Prony Hydrothermal Field was more similar to that of terrestrial serpentinizing sites, while metagenomes from Lost-City were more similar to those from oceanic basalt-hosted vents. One possible explanation is that microorganisms from Lost-City Hydrothermal Field and basalt-hosted vents are subject to numerous physical constraints (hydrostatic pressure) and chemical stresses (heavy metals, radionuclides, etc.), which are probably specific to deep hydrothermal vent and may have more weight in shaping the microbial assemblage. Concerning the specific functional capabilities of microbial inhabitants from serpentinizing sites, this study confirmed the importance of hydrogen-related metabolisms but also revealed a set of highly enriched genes involved in response to environmental stresses. Besides, all genes associated with a pathway of phosphonate catabolism were overrepresented in the metagenomes from serpentinite-hosted ecosystems. We estimated that up to 44% of the microorganisms in these ecosystems possessed the genes encoding for a carbon-phosphorus (C-P) lyase, the key enzyme of this metabolic pathway. These genes were detected in a wide range of bacterial taxa, most of which belonged to *Clostridiales*, *Alpha-* and *Beta-proteobacteria*. Through this pathway, phosphonates can be used by microorganisms to obtain phosphorus in ecosystems where inorganic phosphate is scarce. Moreover, the degradation of phosphonates based on C-P lyase activity lead to hydrocarbons release, such as methane and must be considered as a potential source of biotic methane in the serpentinizing environments.

This study constitutes the first metagenomic comparison of a wide range of serpentinizing systems, providing a better picture of the microbial diversity and the associated gene content. Our results suggest that the phosphonate metabolism, through the C-P lyase pathway, is widespread among serpentinizing systems and is likely to play a role that will require further investigations not only in phosphorus, also in carbon biogeochemical cycles. This metagenomic analysis highlights opportunities for future studies to quantify the reduced phosphorus compounds in serpentinizing environments and check the importance of phosphonate metabolisms using transcriptomic and proteomic analyses.

References

[**1**] Schrenk et al., Reviews in Mineralogy & Geochemistry 75, 575-606 (2013)

[**2**] Russell et al., Geobiology 8, 355-371 (2010)

[**3**] Brazelton et al., Applied and Environmental Microbiology **79**, 3906-3916 (2013)

[**4**] Quéméneur et al., Environmental Microbiology Reports **6**, 665–674 (2014)

Keywords: Comparative metagenomics, Functional capabilities, Random Forest, Serpentinization,

Phosphonate degradation

Assembling the genome of the desert ant and uncovering structural rearrangements with instaGRAAL, a fast and scalable scaffolder based on Hi-C data

Lyam Baudry *† 1, Hugo Darras 2, Martial Marbouty 1, Jean-Francois Flot 2, Serge Aron‡ 2, Romain Koszul§ 1

[1] Régulation spatiale des Génomes - Spatial Regulation of Genomes (RSG) – Centre National de la Recherche Scientifique : UMR3525 – Département Génomes et Génétique - 25-28, rue du Docteur Roux 75724 Paris cedex 15, France

[2] Evolution Biologique et Ecologie – Université Libre de Bruxelles Av. F.D. Roosevelt, 50, CP 160/12 B-1050 Bruxelles, Belgium

Bridging gaps in draft genome assemblies has been a long-standing challenge for a variety of reasons, such as e.g. the presence of repeated sequences. Scaffolding programs often hit an upper bound in the number of contigs they can merge and large eukaryotic genomes are therefore in an unfinished state, very few being scaffolded to a chromosomal level. Here, we present instaGRAAL, a fast, open-source program that uses chromosome conformation capture (Hi-C) data to scaffold contigs based on the collision frequencies between DNA sequences in the nucleus. It uses a simple polymer model to represent the expected spatial contacts between these sequences, and a Markov Chain Monte Carlo method to maximize the likelihood of this model. In order to improve the genome, the program samples each sequence for potential mutations such as insertions, flips or deletions. When applied to the genomes of two hybridogenetic lineages of the desert ant *Cataglyphis hispanica*, instaGRAAL yielded high quality, near-complete, chromosome-level assemblies and uncovered large-scale structural differences that opened new insights into their unusual, hybridogenetic reproductive strategy.

Keywords: whole genome assembly, 3C, chromosome conformation capture, scaffolding, ants, contact genomics, proximity ligation, MCMC

*Speaker

†Corresponding author: lyam.baudry@pasteur.fr

‡Corresponding author: saron@ulb.ac.be

§Corresponding author: romain.koszul@pasteur.fr

CARNAC-LR : Clustering coefficient-based Acquisition of RNA Communities in Long Reads

Camille Marchet *† 1, Lolita Lecompte 1, Corinne Da Silva 2, Corinne Cruaud 2, Jean-Marc Aury 2, Jacques Nicolas 1, Pierre Peterlongo 1

1 Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) – Universite de Rennes 1, Institut National de Recherche en Informatique et en Automatique – Avenue du général LeclercCampus de Beaulieu 35042 RENNES CEDEX, France

2 Genoscope-Centre national de séquençage (GENOSCOPE) – CEA – Genoscope-Centre National de Séquençage 2, rue Gaston Crémieux CP5706 91057 EVRY Cedex, France

Motivation

Lately, long read sequencing technologies, referred to as Third Generation Sequencing, (TGS, Pacific Bioscience [1] and Nanopore [2]) have brought the opportunity to sequence full-length RNA molecules. In doing so they relax the constraint of transcript reconstruction prior to study complete RNA transcripts. By avoiding limitations of previous technologies, [3, 4] and giving access to the trancripts structure, they might contribute to complement and improve transcriptomes studies. This is particularly crucial for non model species where assembly was required. Many biological questions (finding gene signatures for a trait, finding expressed variants...)[5, 6] are classically addressed using transcriptome sequencing. However, this gain in length is at the cost of a computationally challenging error rate (up to 15%) that disqualifies previous short-reads methods. In this work we propose to support the analysis of RNA long read sequencing with a clustering method that works at the gene level. It enables to group transcripts that emerged from a same gene. From the clusters, the expression of each gene is obtained and related transcripts are identified, even when no reference is available.

Problem statement

Within a long reads data set, our goal is to identify for each gene the whole set of reads that come from its expression without the help of a reference genome or transcriptome. This problem can be computationally formalized as a community detection problem, where a community (also referred to as a cluster) is the population of reads coming from a same gene. Communities are densely connected groups of nodes, although there exists no rigorous shared definition. Our application problem is non trivial and specific for three reasons:

1-in eukaryotes, it is common that alternative spliced and transcriptional variants (called isoforms) which differ in exon content occur for a given gene. In this case we want alternative

[*] Speaker

[†] Corresponding author: camille.marchet@irisa.fr

transcripts to be grouped in a same cluster;

2- long reads currently suffer from high error rates and computationally challenging error profiles with a majority of indels errors;

3- all genes are not expressed at the same level in the cell, which leads to an heterogeneous coverage in reads of the different genes, then to communities of different sizes including small ones. This can be a hurdle for community detection.

Previous works

The problem in itself is not new, it dates back before the advent of NGS, with Sanger sequencing and the necessity to cluster ESTs. However these methods were tailored to work with lower scalabilty challenges due to the scarcity of data, and a far less important error rate than with current long reads. The concept of community detection is a natural way of depicting our problem. Due to the ambiguity of the community definition, a plethora of methods have been proposed for their detection. Moreover this problem has appeared in many disciplines, taking many slightly different forms according to the application domain. The first approach that brought an important contribution is an algorithm based on *modularity.* Other methods were then proposed as improvements, in particular methods relaxing the definition of communities as objects that can overlap, such as the Clique Percolation Method (CPM) [7].

Contribution

Roughly speaking, resolution strategies can be classified into two trends according to applications and the community of affiliation: a *graph clustering* strategy based on the search for minimal cuts in these graphs and a *community finding* strategy based on the search for dense subgraphs. Our own approach aims to combine the best of both worlds.
The first approach generally searches for a partition into a fixed number of clusters by deleting a minimum number of links that are supposed to be incorrect in the graph. The second approach frequently uses a *modularity* criterion to measure the link density and decide whether overlapping clusters exist, without a priori regarding the number of clusters. Given that it is difficult to decide on the right number of clusters and to form them solely on the basis of minimizing potentially erroneous links, the main findings and recent developments are based on the community finding strategy and we will focus our review on this approach.
Our algorithm is based on the concept of *clustering coefficient* and we formalize our problem as finding communities such that a community is a connected component in the graph of similarity having a *clustering coefficient* above a fixed cutoff, and such that communities are disjoint sets. An optimal clustering in k communities is a minimal k-cut of the graph, that is, a set of k disjoint subsets of reads such that the set of edges between two different subsets has minimal size. We then implement heuristics that approximate a result of this problem. They are implemented in a tool dubbed CARNAC-LR (Clustering coefficient-based Acquisition of RNA Communities in Long Reads), integrated into a pipeline. The input is a set of long reads and the output is a file with reads indexes grouped in one line per cluster. Our approach is compared to state of the art algorithms to detect communities. We then show its relevance on a real data set issued from mouse brain transcriptome, produced at the Genoscope in the context of ASTER ANR.

Results

We use a real mouse dataset sequenced using a MinION platform at the Genoscope to compare our solution to other algorithms used in the context of biological clustering and demonstrate its is better-suited for transcriptomics long reads.

We build “ground truth” clusters using mapping routine that are compared to *de novo* clustering results.
We use them to benchmark classic community detection algorithms and state of the art sequence clustering tools such as CD-HIT [8] and show we perform better on ONT reads from mouse.
When a reference is available thus mapping possible, we show that it stands as an alternative method that predicts complementary clusters.

References

Manuel L Gonzalez-Garay. Introduction to isoform sequencing using pacific biosciences technology (iso-seq). In Transcriptomics and Gene Regulation, pages 141–160. Springer, 2016.

Jason L Weirather, Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. Comprehensive comparison of pacific biosciences and oxford nanopore technologies and their applications to transcriptome analysis. F1000Research, 6, 2017.

Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data. Nature biotechnology, 29(7):644, 2011.

Tamara Steijger, Josep F Abril, Par G Engstrom, Felix Kokocinski, Tim J Hubbard, Roderic Guigo, Jennifer Harrow, Paul Bertone, RGASP Consortium, et al. Assessment of transcript reconstruction methods for rna-seq. Nature methods, 10(12):1177–1184, 2013.

Robert Ekblom and Juan Galindo. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, 107(1):1, 2011.

Chien-Yueh Lee, Yu-Chiao Chiu, Liang-Bo Wang, Yu-Lun Kuo, Eric Y Chuang, Liang-Chuan Lai, and Mong-Hsun Tsai. Common applications of next-generation sequencing technologies in genomic research. Translational Cancer Research, 2(1):33–45, 2013.

Gergely Palla, Albert-Laszlo Barabasi, and Tamas Vicsek. Quantifying social group evolution. arXiv preprint arXiv:0704.0744, 2007.

Li, Weizhong, and Adam Godzik. ”Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.” *Bioinformatics* 22.13 (2006): 1658-1659.

Keywords: transcriptomics, oxford nanopore technologies, long reads, simulation, mRNA

The genome assembly of a hot ammonia-oxidizing archaeon illuminates the origins of Thaumarchaeota's evolutionary success.

Sophie Abby *† [1,2], Melina Kerou , Christa Schleper

[1] Techniques de lÍngénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications [Grenoble] (TIMC-IMAG) – Centre National de la Recherche Scientifique : UMR5525, Université Grenoble Alpes – Domaine de la Merci - 38706 La Tronche, France
[2] Archaea Ecology and Evolution; University of Vienna, Austria – Austria

Over the past 10 years, Thaumarchaeota have been shown to range among the dominant microbial lineages in various ecosystems. They are found abundantly in the ocean's plankton, but also in deep-sea marine sediments, estuaries and terrestrial environments, making them one of the most successful moderate, aerobic lineage of archaea. Thaumarchaeota are also the only phylum known so far comprising a whole clade which cultured representatives consistently show the ability to oxidize ammonia for energy production. Early diverging clades of Thaumarchaeota are so far represented by species from hot environments, suggesting a thermophilic ancestor for the phylum, and a secondary adaptation to mesophily. We could recently assemble and obtain genomes of un-sequenced representatives within Thaumarchaeota, from environmental samples or enrichment cultures, including the first genome of a thermophilic ammonia-oxidizing archaeon (AOA). Obtaining genomes for these pivotal clades enabled a comparative genomics approach to investigate the origins, and reasons for the evolutionary success of AOA in moderate environments.

Keywords: Evolutionary genomics, Archaea, Adaptation, Phylogenomics

*Speaker

†Corresponding author: sophie.abby@univ-grenoble-alpes.fr

Metagenomic analysis of ancient DNA from dental calculus

Céline Bon * [1], Maxime Borry *

, Marjan Mashkour [2], Julio Bendezu Sarmiento [3]

[1] UMR7206 – Muséum National d'Histoire Naturelle, Département Homme et Environnement, CNRS UMR7206, Musée de l'Homme, 17 place du Trocadéro et du 11 novembre, 75116 Paris – France
[2] Centre National de Recherche Scientifique, CNRS – CNRS : UMR7209 – France
[3] Délégation archéologique française en Afghanistan (DAFA) – Afghanistan

Metagenomic analysis of ancient DNA from dental calculus

Borry, Maxime (1) ; Bendezu-Sarmiento, Julio (2) ; Mashkour, Marjan (3) ; Bon, Céline (1)

(1) Eco-Anthropologie et Ethnobiologie, UMR 7206, Muséum National d'Histoire Naturelle, Centre national de la recherche scientifique, Université Paris Diderot

(2) Délégation archéologique française en Afghanistan

(3) Archéozoologie-archéobotanique : sociétés, pratiques et environnement, UMR7209, Muséum National d'Histoire Naturelle, Centre national de la recherche scientifique

For ten years, the development of next generation sequencing (NGS) has given a boost to ancient DNA (aDNA) studies by opening the field to genome-wide analysis. In a few years, our knowledge about ancient human genetic diversity has been considerably improved (Llamas et al. 2017). However, an increasing number of evidence suggest that our microbiome plays an important role in our evolution and physiology (Turnbaugh et al, 2007). Moreover, our way of living has an impact on our microbiome: ancient microbiome analyses would then provide insights on the ancient cultures and behaviors (Warinner et al. 2014 ; Weyrich et al. 2017). We decided to investigate the content of ancient metagenomes from dental calculus, to better understand diet and microbiome evolution during Protohistory in Southern Central Asia.

However, analyzing ancient microbial data combines challenges of paleogenetics and metagenomics. Because of the effect of time, many post-mortem enzymatic reactions may have happened to the human metagenomics sequences, two of which are the most important: first of all sequences are shortened and fragmented by the action of nucleases and second, sequences can also be affected by deamination, where cytosines lose an amino group and thus become thymines (Briggs et al. 2007).

These two effects combined add up a complexity layer in the analysis of aDNA metagenomics data: while shorter sequences mean less reliable hits when aligning against a reference database,

*Speaker

and shorter contigs when performing de novo assembly, deamination goes with the introduction of mutations in already short sequences that can lead to erroneous taxonomic assignations.

In order to circumvent the inconvenient characteristics of aDNA, in the context of metagenomics studies, we first benchmarked metagenomics tools with simulated (thus controlled) mock communities mimicking the characteristics of aDNA. To do so, we developed ADRSM (Ancient DNA Read Simulator for Metagenomics) which can create mock metagenomic communities mimicking both the short sequences length and the DNA deamination patterns.

To investigate the diet of ancient samples, we developed OrganDiet, a method based on the identification of organelles DNA (chloroplast and mitochondria). Indeed, these multicopy markers have a higher probability to withstand the effect of time, while keeping enough variability to identify clades to the species level. Organdiet was successfully applied to a mock community generated with ADRSM. We compared the performance of Organdiet to other taxonomic assignation methods (Buchfink et al. 2015, Kim et al. 2016), and found that Organdiet performs better, in terms of accuracy and false discovery rate.

We applied Organdiet on an ancient dataset. Analyses of a 65 millions reads dataset from a 13 000 years old cave hyena coprolite samples by Organdiet agrees with results previously obtained on this sample (Bon et al. 2012).

We then applied Organdiet to the 120-130 million reads obtained for each protohistoric Turkmen dental calculus. However, due to the low amount of eukaryotic DNA in these samples, species identification of potential diet elements was not possible.

We then turned to the identification of ancient oral microbiome. Still using an in silico approach with a simulated microbiome mock community, we benchmarked different taxonomic classifier tools on merged reads an assembled contigs, and measured their performances in term of precision and sensitivity. Overall, we found that methods applied on merged reads had a higher sensitivity while methods applied on contigs had a higher precision. Among the best performing methods were MegaBlast (Morgulis et al. 2008) for contigs, and MALT (Herbig et al. 2017), Kraken (Wood et al. 2014), and Metaphlan2 (Truong et al. 2015) for merged reads. As a consequence, we will turn to a combination of these methods for the microbiome analyses of ancient dental calculus.

Overall, we demonstrated that short sequence length remains a bottleneck for metagenomic taxonomic classifiers, and that, depending on the classifier method used, confident species assignation can be tricky to obtain for sequences shorter than 130 bp, which is a problem in ancient DNA.

In order to circumvent the problem of sequence length leading to the impossibility of reliable taxonomic assignation to a specie/genus level, we added a de novo metagenome assembly approach to assemble reads into contigs of greater length.

Using this approach, we managed to increase the precision of most metagenomics methods tested when aligning the contigs, in comparison to the direct alignment of merged reads. Furthermore, de novo assembly of metagenomes is a stepping stone toward a functional analysis, giving us insights on gene-sequences, and thus, functions of the microbial community. Last but not least, metagenome assembly approaches are the only way to recover genomes of species not previously referenced in databases,

In conclusion, to circumvent the effect of DNA degradation in aDNA metagenomics, we advise adding an assembly step together with mapping to increase the precision of aligners and reduce their false discovery rate.

References

Bon, C., Berthonaud, V., Maksud, F., Labadie, K., Poulain, J., Artiguenave, F., et al. (2012). Coprolites as a source of information on the genome and diet of the cave hyena. Proc. R. Soc. B, rspb20120358.

Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Pr´ufer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., et al. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. PNAS 104, 14616–14621.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods 12, 59–60.

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research*, *26*(12), 1721-1729.

Herbig, Alexander, et al. ”MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman.” (2017).

Llamas, B., Willerslev, E., & Orlando, L. (2017). Human evolution: a tale from ancient genomes. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1713), 20150484.

Morgulis, Aleksandr, et al. ”Database indexing for production MegaBLAST searches.” *Bioinformatics* 24.16 (2008): 1757-1764.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods 12, 902–903.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. Nature, 449(7164), 804.

Warinner, C., Rodrigues, J. F. M., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., et al. (2014). Pathogens and host immunity in the ancient human oral cavity. Nature Genetics, 46(4), 336–344.

Weyrich, L. S., Duchene, S., Soubrier, J., Arriola, L., Llamas, B., Breen, J., et al. (2017). Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature, 544(7650), 357–361.
Wood, Derrick E., and Steven L. Salzberg. ”Kraken: ultrafast metagenomic sequence classification using exact alignments.” *Genome biology* 15.3 (2014): R46.

Keywords: ancient dna, metagenomics, microbiome

Gene expression analysis of duck liver steatosis from hybrid and parental species

Xi Liu * [1], Frédéric Hérault * † [2], Christian Diot * ‡ [2], Erwan Corre * § [1]

[1] CNRS, Sorbonne Université, FR2424, ABiMS platform, Station Biologique, 29680, Roscoff, France – Station Biologique de Roscoff – Station Biologique de Roscoff Place Georges Teissier 29680 Roscoff, France

[2] UMR 1348 PEGASE, Physiology, Environment and Genetics for the Animal and livestock Systems, INRA-Agrocampus Ouest, Saint-Gilles, France – INRA-Agrocampus Ouest – UMR1348 PEGASE, 65 rue de Saint-Brieuc, 35044 Rennes Cédex, France

Liver steatosis can occur spontaneously in wild waterfowls as a result of energy storage before migration. This ability is exploited since thousand years in domesticated birds to produce "foie gras" by overfeeding. However, different abilities for fatty liver production are known according to species. Common ducks "Pekin" (Anas platyrhynchos) have a lower ability when compared to Muscovy ducks (Cairina moschata) and to mule ducks (interspecific hybrids produced from a male Muscovy duck and a female common duck), mule ducks representing more than 95% of the foie gras production. To better describe the mechanisms involved in hepatic steatosis development and differences between species and hybrids, next-generation sequencing (NGS) and analyses can be performed on RNAs extracted from the livers of Pekin and Muscovy duck species and of their reciprocal interspecific hybrids, mule and hinny ducks fed ad libitum or overfed. Usually, such RNA-seq analyzes involve the mapping of reads on a reference genome. However, when two different species are involved some mapping biases could be expected. Thus, alternative methods must be developed to be more appropriate for transcriptome analyses and comparisons between the two species and their hybrids.

The whole transcriptomic project includes 80 Illumina paired-end libraries with a read length of 100bp. The raw dataset for each of the 4 species was assembled independently using Trinity, including a reads cleaning process with Trimmomatic then following by an in silico reads normalization step. A global assembly of the 4 de novo transcriptomes was conducted using DRAP for constructing our reference transcriptome. Quality of the reference transcriptome was then validated with an average rate of 85% using pseudoaligner Kallisto and a Busco completeness of 97%. A functional annotation was performed using Trinotate pipelines. Differential expression analysis was conducted using TMM normalization counts in gene level by DEseq2 and edgeR.

Analyses using both reference genome and de novo methods point out a good performance of the de novo method for the different species treatment revealing a new set of genes differentially expressed.

*Speaker

†Corresponding author: frederic.herault@inra.fr

‡Corresponding author: christian.diot@inra.fr

§Corresponding author: erwan.corre@sb-roscoff.fr

Keywords: RNAseq, Interspecies Assembly, Gene expression, Liver steatosis

Développement et optimisation d'un outil de simulation en C++ de données génomiques en populations spatialisées

Thimothee Virgoulay *† [1], Raphael Leblois‡ [2], Alexandre Dehne-Garcia [3], François David Collin [4]

[1] Université de Montpellier (UM) – Master 2 Bioformatique, Connaissances, Données – 163 rue Auguste Broussonnet - 34090 Montpellier, France
[2] Centre de biologie et gestion des populations (CBGP) – Institut national de la recherche agronomique (INRA) : UMR1062 – Campus international de Baillarguet - 34398 Montpellier Cedex 5, France
[3] Centre de Biologie pour la Gestion des Populations (CBGP) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement : UMR55, Centre international d'études supérieures en sciences agronomiques : UMR1062, Institut national de la recherche agronomique [Montpellier] : UMR1062, Université de Montpellier : UMR1062, Institut de Recherche pour le Développement : UMR1062, Institut national d'études supérieures agronomiques de Montpellier – 755 avenue du Campus Agropolis, 34988 Montferrier sur Lez, France
[4] Institut Montpelliérain Alexander Grothendieck (IMAG) – Université de Montpellier, Centre National de la Recherche Scientifique : UMR5149 – UMR CNRS 5149 - Université Montpellier 2, Case courrier 051, 34095 Montpellier cedex 5 - France, France

Le développement rapide des techniques de séquençage, ainsi que des moyens de calculs informatiques a révolutionné la génétique/génomique des populations ces 20 dernières années. Cependant, du fait de la difficulté à obtenir des données génomiques individuelles et géo-référencées pour de gros échantillons et de la lourdeur des modèles spatialement explicites, l'aspect spatial a souvent été délaissé.

On assiste aujourd'hui à deux changements majeurs : les génomes individuels atteignent un prix raisonnable, les méthodes d'estimations basées sur la simulation de type ABC (Approximate Bayesian Computations) ont gagné un facteur 10 à 100 en terme d'efficacité grâce à l'utilisation des algorithmes de forêts aléatoires (Pudlo et al. 2015).

De ce fait, il est maintenant envisageable de faire de l'estimation de paramètres démographiques (dispersion, densité et tailles de populations) et historiques (dater des changements démographiques) à partir de données génomiques sous des modèles démo-génétiques spatialisés de plus en plus réalistes.
Améliorer ces techniques d'inférences et les modèles sous-jacents permettra de répondre à des questions essentielles pour mieux comprendre la répartition et l'évolution de la diversité génétique des populations dans le temps et l'espace.

*Speaker

†Corresponding author: thimothee.virgoulay@etu.umontpellier.fr

‡Corresponding author: raphael.leblois@supagro.inra.fr

Notre but est d'implémenter un nouveau simulateur de données génomiques basé sur des algorithmes de coalescences pouvant considérer des modèles spatialisés, afin de l'utiliser pour faire de l'inférence démo-génétique. Les techniques modernes d'inférence, par ABC entre autres, nécessitant des algorithmes efficaces, autant en terme de vitesse d'exécution des calculs que de l'espace mémoire nécessaire, le choix des méthodes de stockage et d'indexation des arbres de coalescence et des génomes simulés est donc crucial pour permettre de simuler de gros jeux de données très rapidement (Kelleher et al. 2016).

Ce projet vise le développement d'un logiciel autonome, open source, collaboratif (Git) et si possible en intégration continue. Il est organisé de façon à s'orienter vers une programmation dite " moderne " en C++, utilisant de manière extensive les nouveautés du standard (C++11/14/17), de manière à produire un code lisible, concis, optimisé et immédiatement réutilisable.

Keywords: c++, génétique des populations, modèles spatiaux, simulateur

Discovering Millions of Plankton Genomic Markers from the Atlantic Ocean and Mediterranean Sea

Majda Arif * [1]

[1] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France – CEA, Genoscope – France

The comparison of the molecular diversity in all plankton populations present in geographically distant water columns allows in theory a holistic view of the connectivity, isolation and adaptation of organisms in the marine environment. In this context, the large-scale analysis of genomic variants in metagenomic data appeared to become a powerful strategy to detect new makers that enable the identification of genetic structures and genes under natural selection in plankton.

Here, we used *DiscoSnp++,* a reference-free variant caller to produce genetic variants from large-scale metagenomic data and tested its accuracy in terms of variant calling, allele frequency and population genomic statistics by comparing to the state-of-the-art method. *DiscoSnp++* produces less false positive variants, more accurate allele frequencies, similar genetic structure and identification of loci under natural selection. *DiscoSnp++* was then applied to 114 metagenomic samples from four fraction sizes ranging prokaryotes, protists and zooplankton sampled from 35 *Tara* Oceans sampling stations located in the Atlantic Ocean and Mediterranean Sea. We produced a new set of marine genomic markers containing more than 19.106 variants.

This new genomic resource can be used by the scientific community to relocate these markers on their genomes and transcriptomes of interest and can take advantage of this universal resource that will be updated with new marine expeditions and the increase of metagenomic data production (availability: http://bioinformatique.rennes.inria.fr/taravariants/)

Keywords: Tara Oceans, Population genomic, Selection, Oithona, DiscoSnp

*Speaker

Introducing Metavariant Species for Reference-Free Population Genomics Analysis using Metagenomics

Laso-Jadart Romuald *† [1], Majda Arif [1], Patrick Wincker [1], Pierre Peterlongo [2], Mohammed-Amin Madoui‡ [1]

[1] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay – CEA, CNRS, Université Paris-Saclay – Evry, France

[2] IRISA Inria Rennes Bretagne Atlantique, GenScale team – INRIA-IRISA – France

The availability of large datasets of metagenomic sequences offers new opportunities for population genetic research. However, for many non-model species, the lack of reference genomes remains an important issue and constitutes an obstacle for population genomic studies. Here, we developed a new reference-free method to identify species named metavariants species (MVS) by analogy to the metagenomic species (MGS). After detecting biallelic loci directly from metagenomic reads using discoSnp++, MVSs are identified by density-based clustering on biallelic loci depth sequencing coverage across all sampled populations. Then, the allele frequencies of MVS can be used for population genomic analyses to identify population differentiation and loci under natural selection. We applied this method to decipher population structure and differentiation on *Tara* Oceans metagenomic data generated from the Mediterranean Sea plankton. For some MVS, we found a correlation between the genetic differentiation and environmental parameters. We were also able to detect loci under local adaptation in each MVS, showing that classical population genomics analyses are possible with this new method.

Keywords: Metavariant species, population genomics, metagenomics : Tara Oceans

*Speaker

†Corresponding author: rlasojad@genoscope.cns.fr

‡Corresponding author: amadoui@genoscope.cns.fr

Bacterial genomes assembly using PacBio long read technologies

Sophia Achaibou * [1], Thomas Cokelaer * † [2]

[1] Institut Pasteur - Biomics Pôle - CITECH – Institut Pasteur de Paris – 25-28, rue du docteur Roux, 75724 Paris cedex 15, France

[2] Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 26-28 Rue du Docteur Roux, 75015 Paris, France

Even though bacterial genomes are small (as compared to eukaryotes), they may be difficult to assemble due to their high GC content or large repeated regions. Assembly of such genomes remains a challenging task especially with short-read technologies (Illumina). For instance, the Bordetella bacteria has a GC content of 68% and about 20% of the genome is made of repeated regions. In addition to known biases due to high GC content, the short Illumina reads (_˜100 bases) do not cover long repeated regions (e.g., IS regions are 1,000 bases long) making assembly a difficult or impossible task. On the contrary long reads technologies like PacBio can avoid such problems. First, because there is no bias with respect to GC content but also because the mean length of the reads (15,000 bases) can cover highly repeated regions. Here, we present results regarding the assembly of bacterial genomes using PacBio sequencing data that have high GC content and large repeated regions. We show that it can be quite easy to build an assembly from raw pacbio data. We show what should be the minimal coverage to obtain a single contig and what quality can be obtained with such coverage in terms of mismatches, indels, and accuracy. More particularly, we present the assembly results for a wide range of coverage on several species including Bordetella pertussis and Veillonella, which are known for having large GC content and repeated regions. The results can be used to reduce the cost of PacBio sequencing by targeting a specific coverage and also to perform multiplexing.

Keywords: PacBio, assembly, repeated regions, high GC content

*Speaker

†Corresponding author: thomas.cokelaer@pasteur.fr

Data Integration and VISualization (DIVIS) : from large heterogeneous datasets to interpretable visualisations in plant science

Ophélie Thierry * [1], Rachid Boumaza [1], Julia Buitink [1], Claudine Landès * † [1], Olivier Leprince [1], Mathilde Orsel [1], Pierre Santagostini [1], Julie Bourbeillon‡ [1]

[1] Institut de Recherche en Horticulture et Semences (IRHS) – Université d'Angers, Institut National de la Recherche Agronomique : UMR1345, Agrocampus Ouest, SFR 4207 QuaSaV – 49071, Beaucouzé, France

The demand by biologists to integrate heterogeneous and large datasets from "omics" and phenotyping activites is rapidly increasing [1, 2, 3]. However, methods automating this approach are still at its infancy and to our knowledge, no operational and user-friendly software yet exists. Experiments are performed in- dependently and resulting data are cross-analysed manually and a-posteriori by scientists [4]. For instance, the biology teams from the IRHS (Institut de Recherche en Horticulture et Semences) in Angers have been accumulating datasets of different natures (transcriptomic, biochemistry, physical measures, sensory analysis, etc.) regarding perennial, annual and biannual plants. These datasets are described using reference ontologies enriched with in-house knowledge and stored in a Laboratory Information Management System (LIMS) which is developed and distributed by the IRHS Bioinformatics team.
The main objective of the DIVIS (Data Integration and VISualization) project is to develop a directly usable prototype of a new data analysis tool, by combining the most promising integration and visualisation approaches that are publicly available using the heterogeneous large scale datasets stored in our LIMS.

As a first step, the tool will download and normalise experimental datasets in respect with samples of similar nature across different scales ranging from the molecule to the organism, types of experiments and experimental designs, which is seldom performed by existing software, in particular in plant biology. The originality of our new integration approach, features the following analysis of the resulting matrix:

1. reduce the number of individuals by regrouping similar samples using a similarity score,

2. calculate this score based on similarity between metadata variables stored in a specifically designed ontology. For each ontology concept, relationships with its neighbours will be associated with similarity indices. These indices will be used to calculate a similarity between individuals associated with these ontology concepts [5],

*Speaker

†Corresponding author: claudine.landes@inra.fr

‡Corresponding author: julie.bourbeillon@agrocampus-ouest.fr

3. represent each group by an archetype sample,

4. construct graphical representations of the results. The visualisation approach will allow to present data regarding these archetype samples in a multi-layer display separating various subsets of coherent data and to navigate through the results.

In order to validate the methodology and assess how the approach can be adapted to different experimental contexts with an equivalent level of complexity, the tool is developed based on two test datasets acquired as part of matching experiments (including several studies performed on the same samples):

• an apple fruit dataset including descriptors at the variety level (fruit shape, colour or size, tree shape or vigour, etc.) and measures at the fruit level (transcriptomic, biochemical, physical and sensory data),

• a seed dataset containing descriptors at the genotype level (genotype, environmental and climatic data regarding the collection site) and at the seed level (germination kinetics and physical attributes).

So far we have constructed the integrated and normalised data matrices. We are currently designing or reusing relevant ontologies [1, 5, 6, 7] and associating each individual with concepts from these ontologies. The aspects that are under consideration are as follows:

• For the apple dataset:

– reuse existing lists of descriptors associated with apple varieties (fruit colour, fruit shape, etc.) to devise ontologie.

• For the seed dataset:

– design an ontology of the shape (long, short, straight, curved, etc.) of seeds and associate these concepts with actual measurement,

– reuse exisiting colour tables and add relationships between colours to devise a colour ontology associated to HSV values to describe seeds colour,

– reuse the K$\acute{}$oppen climate classification to associate climatic data for each genotype collection site to a climate clas,

– reuse existing lists of descriptors associated with seed collection sites (topology, pedology, etc.) and plant characteristics for each genotypes (pod shapes, flower colours, leaf shapes, etc.) to devise ontologies,

The next stage will be to cluster the individuals according to these ontology concepts.

References

Solovieva E, Shikanai T, Fujita N, Narimatsu H. GGDonto ontology as a knowledge-base for genetic diseases and disorders of glycan metabolism and their causative genes. Journal of Biomedical Semantics. 2018;9:14. doi:10.1186/s13326-018-0182-0.

Hendler J. Data Integration for Heterogenous Datasets. Big Data. 2014;2(4):205-215. doi:10.1089/big.2014.0068

Fonseca, Frederico T. et al. ”Using Ontologies for Integrated Geographic Information Systems.” Trans. GIS 6 (2002): 231-257.

Arguello Casteleiro M, Demetriou G, Read W, et al. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. Journal of Biomedical Semantics. 2018;9:13. doi:10.1186/s13326-018-0181-1.

Kohler S. Improved ontology-based similarity calculations using a study-wise annotation model. Database: The Journal of Biological Databases and Curation. 2018;2018:bay026. doi:10.1093/database/bay026.

Cohen J, Matthen M, Bradford Book, A. (2010). Color Ontology and Color Science.

Hartmann, J, Palma, R, Gómez-Pérez, A. (2009). Ontology Repositories. Handbook on Ontologies.

Keywords: Heterogeneous Datasets, Data Integration, Ontology, Plant Science

Detection and Characterization of Genomic Variations using Running Median and Mixture Models: application to CNV detection.

Thomas Cokelaer *† [1], Dimitri Desvillechabrol

[1] Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 26-28 Rue du Docteur Roux, 75015 Paris, France

Sequencing technologies allow researchers to investigate a wide range of genomic questions, covering research fields such as the expression of genes (transcriptomics), the discovery of somatic mutations, or the sequencing of complete genomes of cancer samples to name a few examples. The emergence of the second generation sequencing, which is also known as Next-Generation Sequencing or NGS hereafter, has dramatically reduced the sequencing cost. This breakthrough multiplied the number of genomic analyses undertaken by research laboratories but also yielded vast amount of data. Consequently, NGS analysis pipelines require efficient algorithms and scalable visualization tools to process this data and to interpret the results.

Raw data generated by NGS experiments are usually stored in the form of sequencing reads (hereafter simply called reads). A read stores the information about a DNA fragment and also an error probability vector for each base. Read lengths vary from 35-300 bases for current short-read approaches to several tens of thousands of bases possible with long-read technologies such as Pacific Biosciences or Oxford Nanopore.

After trimming steps (quality, adapter removal), most high-throughput sequencing (HTS) experiments will require mapping the reads onto a genome of reference. If no reference is available, a de-novo genome assembly can be performed. In both cases, reads can be mapped back on the reference taking into account their quality. We define the genome coverage as the number of reads mapped to a specific position within the reference genome.

The theoretical distribution of the genome coverage has been thoroughly studied following the seminal work of Lander-Waterman model. A common metric used to characterize the genome coverage is the sequencing depth: the empirical average of the genome coverage. It may also be called depth of coverage (DOC). The sequencing depth unit is denoted X.

The required sequencing depth depends on the experimental application. For instance, to detect human genome mutations, single-nucleotide polymorphisms (SNPs), and rearrangements, a 30 to 50_˜X depth is recommended in order to distinguish between sequencing errors and true SNPs. In contrast, the detection of rarely expressed genes in transcriptomics experiments often

*Speaker

†Corresponding author: thomas.cokelaer@pasteur.fr

requires greater sequencing depth.

The Lander-Waterman model provides a good theoretical estimate of the required sequencing depth to guarantee that all nucleotides are covered at least N times.

This is, however, a theoretical estimate that does not take into account technical and biological limitations; some regions being difficult to efficiently map (e.g., repetitive DNA) or containing compositional biases (e.g., GC bias). Furthermore, the genome coverage itself may contain a non-constant trend along the genome due to the impact of replication from the origin of replication. Finally, some regions may be deleted or duplicated.

While the sequencing depth and other metrics provides a quick understanding about the quality of sequencing and mapping, the genome coverage can also be analysed to identify genomic variations such as single nucleotide variations (SNVs) or copy number variations (CNVs).

In order to detect genomic regions of interests (ROIs) based on genome coverage, a simple and fast approach might be to set two arbitrary thresholds bounding the sequencing depth. However, there are two major drawbacks with this approach. First, with a fixed threshold, one may detect numerous false signals (type I errors) or fail to detect real events (type II errors). An adaptive thresholds that follows the trend of the genome coverage is thus required. Furthermore, a fixed threshold is arbitrary and so the detected events lack a robust means of assigning significance. A more robust alternative is to estimate the genome coverage profile histogram from which a z-score statistics can be used to identify outliers more precisely. Due to a number of known and unknown biases, one should still normalize the data. There are a number of different methods for detecting the ROIs. For example, for CNV detection, numerous techniques are used such as the mean-shift technique (CNVnator tool) or bias correction followed by application of a complex statistical model (CNOGpro tool).

In this work we describe a novel approach that can efficiently detect various types of genomic ROIs. The algorithm does not target any specific type of genomic variations but instead systematically reports all positions (with a z-score) that have depth departing from the overall distribution. The algorithm normalizes the genome coverage using a running median and then calculate a robust statistic (z-score) for each base position based on the parameter estimation of the underlying distribution. This allows us to obtain robust and non-constant thresholds at each genome position. Various types of clustering or filtering can then be implemented to focus on specific categories of variations.

In the first part of the talk, we describe the proposed novel method of detecting ROIs in the genome coverage data. In particular, we describe (i) the running median used to detrend the genome coverage, (ii) the statistical methods used to characterize the central distribution from which outliers can be identified and (iii) a double thresholds method proposed to cluster the ROIs.

In the second part, we present an application for CNV detections. In particular, in the context of bacterial genomes, we show how this implementation out-performs some established tools in not only detecting CNVs but also precisely identifies their location and number. As a test example, we use 6 isolates of Staphylococcus aureus. We describe the different between our implementation and two established tools namely CNVnator and CNOGpro, the latter being dedicated to the detection of CNV in bacterial genomes.

Reference : https://www.biorxiv.org/content/early/2016/12/08/092478

Keywords: genome coverage, sequencing depth, running median, CNV detection

TRANS-C3 - The Transcriptome of downy oak (Quercus pubescens Wild) in Response to Climate Change

Xi Liu * [1], Beatrice Loriod * † [2], Stefano Caffarri * ‡ [3], Erwan Corre * § [1], Jean-Philippe Mevy * ¶ [4]

[1] CNRS, Sorbonne Université, FR2424, ABiMS platform, Station Biologique, 29680, Roscoff, France – Station Biologique de Roscoff – Station Biologique de Roscoff Place Georges Teissier 29680 Roscoff, France

[2] Transcriptomique et Genomique Marseille-Luminy (TGML) – TAGC U1090 Inserm – Campus Universitaire de Luminy 13009 Marseille. TAGC U1090 Inserm, Aix Marseille Universite, France

[3] Biologie cellulaire et moléculaire et des plantes et bactéries, Institut de Biologie Environnementale et Biotechnologie (IBEB), Laboratoire de génétique et biophysique des plantes (LGBP) – CEA-CNRS, UMR 6191 – Faculté des Sciences de Luminy, Case 901, 163 av de Luminy13288 Marseille cedex 9, France

[4] IMBE - UMR CNRS 7263 / IRD 237 Equipe Diversité et Fonctionnement : des Molécules aux Ecosystèmes(DFME) – UMR CNRS 7263 – Aix-Marseille Université, Centre Saint-Charles - Case 4, France

The TRANS-C3 project fits into the context of global change and aimed to understand the response of forest trees to the aridification of the Mediterranean climate. For this purpose, researches have been conducted using downy oak model at the O3HP site, where the drought prediction is simulated by a reduction in rainfall amount to about 30%. Precisely the objective was to understand *in situ* plant response to climat change both in terms of gene expression, metabolic footprint and the impact on an essential physiological process: photosynthesis.

To adress the question of the identification of candidate genes involved in the response to drought, the analyses were carried out using leaf samples from 20 trees (10 in rain exclusion and 10 in control) harvested over 2 periods (Spring-Summer), the whole transcriptome project includes 40 samples. RNA sequencing was performed by Next-Seq Illumina 500 for paired-end libraries with a read length of 150 bp which allowed to generate 1 971 431 570 x 2 raw reads. Based on the genome of *Quercus robur* constructed by INRA in Bordeaux, we assembled a *Quercus pubescens* transcriptome using genome-guided Trinity *De novo* transcriptome assembly protocol. The raw assembly corresponds to 530080 transcripts (and 395969 "Trinity genes"), reduced to 156986 highly expressed transcripts (TPM > 1 and isoforms> 1%) (and 126106 "Trinity genes"). The quality of the reference transcriptome was then validated with an average remapping rate of 86% using pseudoaligner Kallisto and a Busco completeness of 88%. About 30% of the transcripts could be annotated by the Trinotate pipeline. Differential expression analysis (DESeq2/edgeR) allows to identify 42 transcripts involved in the adaptation of downy oak to climate aridification

*Speaker

†Corresponding author: beatrice.loriod@inserm.fr

‡Corresponding author: Stefano.caffarri@univ-amu.fr

§Corresponding author: erwan.corre@sb-roscoff.fr

¶Corresponding author: jean-philippe.mevy@imbe.fr with a very pronounced seasonal effect. Similarly, the metabolic study targeted on metabolites of the tricarboxylic acid pathway, sugars and amino acids has identified chemical markers of the effect of aridity such as pyroglutamic acid, and markers of the seasonal effect as xylulose and sorbitol.

We also tested the response of *Q. pubescens* to aridification in terms of the plasticity of the photosynthetic machinery. Three methods were used: (i) by western blot to check photosystems protein modifications; (ii) by HPLC for the pigment composition (chlorophylls and carotenoids) and (iii) by fluorimetry through *in situ* monitoring of the response to variations in light intensity. Overall, except for PsbS, the results indicate that in the spring plants are more stressed than summer with a very small difference between control and exclusion for each season. Similarly, the kinetics of quantum yields (ϕPSII) have identical profiles regardless of the treatment (control/exclusion) and the season. On the other hand, quenching mechanisms differ mainly in the summer.

In conclusion: This project has identified genetic, biochemical, chemical and physiological markers of plant adaptation to climate change in a cross-cutting approach. It has been carried out around several disciplinary fields: Genomics, Biochemistry, Ecophysiology, Chemistry, Bioinformatics and Modeling, opening perspectives for more ambitious projects.

Keywords: RNAseq, Genome guided de novo transcriptome assembly, Gene expression, Oak, Global change, Cross cutting approach, Photosynthesis, Metabolomics

PiRATE v2: detection, classification and annotation of transposable elements of non-model organisms

Jérémy Berthelier *† 1, Tristan Frances 2, Myriam Badawi 2, Véronique Jamilloux 3, Nathalie Casse 2, Bruno Saint-Jean 1, Grégory Carrier 1

1 Physiology and Biotechnology of Algae Laboratory (PBA) – Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER), Nantes – Rue de l'Ile d'Yeu BP 21105 Nantes cedex 3, France
2 Laboratoire Mer, Molécules, Santé (EA 2160) (MMS) – Université du Maine – Avenue Olivier Messiaen 72085 Le Mans cedex 09, France
3 URGI INRA de Versailles – Institut National de la Recherche Agronomique - INRA – France

Transposable elements (TEs) form a group of diverse mobile DNA sequences, which constitute powerful forces of the genome evolution of eukaryotic organisms. Their annotation in genome of non-model organisms is challenging because their genome assembly is usually not at chromosomal level and is fragmented, increasing the difficulty to detect them. Moreover, non-model organisms usually belong to poorly studied taxa, where few TEs have been characterized, increasing the difficulty to recognize/classify them with basic similarity methods.
To counter these limitations, we designed a bioinformatics pipeline named PiRATE (Pipeline to Retrieve and Annotate TEs) to detect, classify and annotate TEs of non-model organisms. PiRATE combines multiple analysis packages, representing all the major approaches for TE detection. The goal of PiRATE is to promote the detection of full-length TE sequences to facilitate their classification. The classified sequences are then used as a *de novo* TE library to perform the TE annotation.

Here, we present a second version of PiRATE. While its classification step was previously performed solely with the tool PASTEC (Hoede et al, 2014), which classify the detected sequences at the TE Order level (e.g. LTR or TIR), this second version includes a complementary script allowing automatic classification of TEs at the superfamily level (e.g. LTR/Copia or TIR/Mariner). This improvement facilitates the clustering of classified TE sequences into families and greatly decreases the manual step.
PiRATE was controlled with genomic data of the model plant *Arabidopsis thaliana* and is able to detect 81% of its TE families. We also applied it with the genome of the non-model species *Tisochrysis lutea* and discovered that it is composed of 3.79% and 17.05% of potentially autonomous and non-autonomous TEs, respectively. PiRATE is automated in a stand-alone Galaxy and is available through a virtual machine: http://doi.org/10.17882/51795

*Speaker
†Corresponding author: jeremy.berthelier@ifremer.fr

Keywords: Genomics, Annotation, Tranposable elements, Pipeline

Metabarcoding on the deep seafloor: optimizing multigene approaches for large-scale biodiversity assessments

Caroline Dussart * [1,2], Sophie Arnaud-Haond† [3,4], Miriam Brandt‡ [4]

[1] Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) – Ministère de l'Enseignement Supérieur et de la Recherche Scientifique – France
[2] Université de Rouen Normandie (UNIROUEN) – Ministère de l'Enseignement Supérieur et de la Recherche Scientifique – 1, rue Thomas Becket 76821 Mont-Saint-Aignan Cedex, France
[3] Observatoire de REcherche Méditerranéen de l'Environnement (OSU OREME) – CNRS : UMS3282, Institut de recherche pour le développement [IRD] : UMS223, Université Montpellier II - Sciences et techniques – UM2, Bât. 22 - CC 060, Place Eugène Bataillon - 34095 Montpellier Cedex 5, France
[4] MARine Biodiversity, Exploitation and Conservation (MARBEC) – UMR 9190, Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) – Avenue Jean Monnet, CS 30171, 34203 Sete Cedex, France, France

The deep sea, the largest and most poorly known biome on Earth, is under increasing threat from human-induced ecological impacts. Improved baseline knowledge and environmental impact assessment protocols are required to be able to alleviate potential changes in ecosystem diversity and functioning in the deep sea. Metabarcoding of environmental DNA (eDNA) enables broader and faster biodiversity assessments, and is increasingly used to study eukaryote and prokaryote diversity. Whether metabarcoding provides reliable diversity inventories that meet the quality standards for accurate baseline data and biomonitoring is still uncertain in the deep-sea benthos, the latter being associated with specific taxonomic and sampling challenges.

Before launching a large-scale project for the reassessment of deep-sea biodiversity, we addressed these technical challenges using bathyal and abyssal sediments sampled in the Mediterranean and central Atlantic, aiming to setup optimized protocols and evaluate the strengths and limitations of multigene metabarcoding in the deep sea. Five mock communities, whose composition in terms of taxa and abundance were precisely known, were used to assess whether we could find satisfying biodiversity assessments with barcode genes 18SV1V2, COI and 16SV4V5. Using these mock communities, we combined several tools into a pipeline as follows (work in progress):

1) Cleaning of the reads:

- primer removal

- filtering of poor-quality reads with R package DADA2 [1]. The filtering method allows to remove reads with too many expected errors (obtained by summing the risk of each base to be an error). We currently remove reads with more than 2 expected errors on forward reads and 4

*Speaker

†Corresponding author: sophie.arnaud@ifremer.fr

‡Corresponding author: miriam.brandt@ifremer.fr

errors on reverse reads.

- read correction with DADA2. We used to obtain a surprisingly high demultiplication of OTUs on our mock communities, revealing that sequencing errors were numerous even though using Illumina data. Using DADA2 significantly reduced the demultiplication of OTUs on our mocks dataset. The DADA2 algorithm makes use of a parametric error model, producing a matrix of error rates (A-> C, A-> G, etc.). The error model is learnt from the data, using the quality scores of the reads, by alternating estimation of the error rates and inference of sample composition until they converge on a jointly consistent solution.

- merging of forward and reverse reads with DADA2. As DADA2 corrected the reads, they recommend to allow 0 mismatch for the merging. However, we chose to allow up to 2 mismatches, to make sure we don't lose too many reads that may have not been completely corrected.

- filtering of sequence variants that are either too long or too short. These length must take into account the natural variation of the genes. We currently apply a (300-500) filtering to all the studied loci, as we expect 313bp for COI, 411bp for 16S, and 450bp for 18S (but actually rather get 370).

- chimera removal with DADA2. Bimeras are identified by performing a Needleman-Wunsch global alignment of each sequence to all more abundant sequences, and then searching for combinations of a left-parent and a right-parent that cover the child sequence without any mismatches or internal indels. Mismatches and internal indels are not allowed, because DADA2 has already corrected the sequences. Variants are therefore supposed to be real sequence variants.

2) Clustering of the sequence variants into OTUs

We compared two different ways of clustering:

- similarity threshold based algorithms : UCLUST [2]

- nucleotide distance based algorithm : SWARM [3]

Similarity threshold based algorithms have arbitrary global clustering thresholds (96/97% work well for most taxa, but may need a refinement for some others) and are input-order dependent, while SWARM doesn't use similarity thresholds and provides consistent results. We compared the results of these clustering methods on our mock communities and didn't find much difference between them in terms of species that were found. SWARM results were slightly better, though.

3) Taxonomical assignment

The sequence variants are assigned with blastn [4], against:

- 16S and 18S: Silva 132 [5]

- COI : Midori unique [6]. We may also use GenBank for COI when not getting enough results from Midori (which is however more curated).

On our mock communities, most OTUs were not assigned to the species we were looking for, but to close species. This was expected, as the barcode genes we used are known not to be resolutive to the species level. A small number of them were also completely wrongly assigned.

When working on real datasets from the deep sea, the main challenge with the taxonomical assignment will be that deep-sea species are poorly referenced, which will probably lead to many unassigned OTUs.

4) Removal of spurious OTUs

Spurious OTUs are removed with LULU R package [7]. LULU identifies errors by combining sequence similarity and co-occurrence patterns. This allows to decrease the demultiplication of OTUs due to sequencing errors. Spurious OTUs are very few after the DADA2 correction of the reads, though.

5) Removal of numts

The numts (nuclear mitochondrial DNA) are nuclear copies of the genes we use. They evolve differently (no selection) and we are therefore not interested in them. However, we haven't found yet a satisfactory way to remove them from the dataset.

6) Bio analyses

The bio analyses will be conducted with Phyloseq R package [8] : alpha and beta diversity.

Conclusion

The pipeline we created could find about 100% of the eukaryotic OTUs we knew were present in our mock communities, provided we used 2 different barcode genes : COI and 18S. We couldn't expect one single barcode to find them all, as barcode genes can't amplify every taxon, even though their primers are designed to be universal. Combining two barcodes allows to better cover tha taxa present in the samples.

We initially had many artefactual OTUs caused by sequencing errors, but the use of DADA2 and LULU efficiently reduced the number of OTUS, leading to a number of OTUs very close to the quantity we were expecting.

The main problem that remains is the taxonomical assignment, with most OTUs getting the correct family and an incorrect genus and species, and a few OTUs getting a completely wrong assignment. This issue hasn't yet been solved, and will be even more problematic when working on real datasets from the deep sea, as most of the organisms living there haven't been referenced yet.

Perspectives:

- apply MACSE to the sequence variants (before clustering them into OTUs), to remove the sequences that contain frameshifts and/or stop codons, so that most numts are removed [9]

- implement custom databases for taxonomical assignment

- test and adapt the pipeline for prokaryotes

References:

1. Callahan,B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S.P. Holmes. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13:581-+

2. Edgar,RC (2010) Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26(19), 2460-2461
3. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2015) Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ 3:e142
4. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-41
5. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Gl´ockner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools
6. Ryuji J. Machida, Matthieu Leray, Shian-Lei Ho & Nancy Knowlton (2017), Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples
7. Tobias Guldberg Frøslev, Rasmus Kjøller, Hans Henrik Bruun, Rasmus Ejrnæs, Ane Kirstine Brunbjerg,Carlotta Pietroni & Anders Johannes Hansen (2017), Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates
8. McMurdie and Holmes (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE. 8(4):e61217
9. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. Vincent Ranwez, Sébastien Harispe, Frédéric Delsuc, Emmanuel JP Douzery. PLoS One 2011, 6(9): e22594.

Keywords: metabarcoding, deep sea, protocols

Building a Metatranscriptomics Strategy to explore Insect Vectors in Africa

Vanesa Assele Koghe *† [1,2], Thomas Cokelaer [3], Hugo Varet [4], Nicolas Berthet [2], Sean Kennedy [1], Catherine Dauga * ‡ [1]

[1] Biomics - CITECH – Institut Pasteur de Paris – 28 rue du Dr Roux, 75724 Paris cedex 15, France
[2] Centre international de recherches médicales de Franceville (CIRMF) – B.P. 769 Franceville, Gabon
[3] Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 26-28 Rue du Docteur Roux, 75015 Paris, France
[4] Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 25-28, rue du Docteur Roux 75724 Paris Cedex 15, France

Environmental Metatranscriptomics aims at estimating the diversity and functions of microorganisms from various environments by direct sampling of culturable and uncultured microorganisms. Many tools already exist to analyze metatranscriptomes but they have never been evaluated in the context of insect studies.
Insect metatranscriptomes are complex mixtures which contain high levels of insect related transcripts and relatively rare microorganism transcripts. Therefore, in the process of establishing an automated analysis pipeline for such cases, it is first necessary to define a fast and efficient strategy to extract information on microorganism transcripts.

Here, we studied metatranscriptomes from insects collected from caves in Gabon. Insects coexist with bats and rodents in these caves and are known to be reservoirs of pathogens. We wanted to determine if insect vectors, living in environment far from human activities, can act as potential vectors of pathogens.

We analyzed a sample of 48 million reads (10 million after trimming) using a classical procedure of metranscriptome analysis, with and without subtraction of insect-derived transcripts. We tested various strategies including three methods of insect-transcript subtraction and two parameters for read assembly. Results were validated by simulation tests and by non-parametric statistical tests appropriate for the non-normal distribution of our sample data.

Results showed that direct exploration of sample biodiversity, without subtraction, yielded non-reproducible results and contained many false negatives. This poor performance was irrespective of the assembly parameters. Subtraction of insect transcripts greatly improved the quality of assembling and led to highly consistent results. We thus demonstrate that subtraction of insect transcripts is an essential step to improve the quality of metranscriptomics data analysis.
In addition, first evidence of active pathogens, such as Loa Loa (filariasis), *Leishmania* (cutaneous leishmaniasis), *Schistosoma* (bilharzia), *Trypanosomia* (sleeping thickness) and *Listeria*

[*] Speaker
[†] Corresponding author: Vanesa.ASSELEKOGHE@pasteur.fr
[‡] Corresponding author: catherine.dauga@pasteur.fr

monocytogenes (Listeriosis), in insects from caverns in Gabon, suggested that future metatranscriptomics studies will be useful for epidemiological surveys of insect vectors in Africa.

Keywords: Metatranscriptomics, Insect vector diseases, Bioinformatics Strategy

Assessing the datation of gene duplication in vertebrates and its link with species diversification

Guillaume Louvel *† [1], Hugues Roest-Crollius [2]

[1] Institut de biologie de l ENS Paris (UMR 8197/1024) (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris, Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique - CNRS : UMR8197 – 46 rue d'Ulm 75005 Paris, France

[2] Institut de biologie de lÉNS Paris (UMR 8197/1024) (IBENS) – École normale supérieure - Paris, Institut National de la Santé et de la Recherche Médicale : U1024, Centre National de la Recherche Scientifique : UMR8197 – 46, Rue d'Ulm75005 Paris, France

Genes within genomes do not necessarily share the same phylogenetic history as species. This discrepancy originates from gene duplications, deletion, horizontal transfer or incomplete lineage sorting. Evolutionary biologists have long been interested in gene duplication, as new gene copies provide the potential for adaptive novelty (Ohno 1970). As a growing number of complete annotated genomes are now available, especially in vertebrates, it becomes possible to quantify the genome-wide history and dynamics of gene duplication (Blomme et al. 2006). A fine-grained temporal distribution of duplications along lineages would enable testing hypotheses about the evolutionary consequences of gene duplication: do they provide key innovations that allow clades to successfully radiate? Do they generate genetic incompatibilities that foster speciation?

To answer these questions, we developped a pipeline to date each gene duplication in absolute time based on synonymous substitution rates (dS) and reconciled gene trees, using existing tools (PAML's codeml for dS (Yang 2007), TreeBest for reconciliation with the species tree (Vilella 2009)). We applied different methods of datation: Mean-Path-Length (Britton et al. 2002), Penalized-Likelihood (Sanderson 2002) and the computationally more intensive DLRS (Sj́ostrand et al. 2012). We then assess the relevance of the inferred duplication dates, and their susceptibility to methodological pitfalls such as robustness of the molecular clock assumption or topology uncertainty in large gene trees.

This analysis allows us to assess the reliability of current methods and data used to perform molecular absolute datation. Accurate gene duplication datations would help us discriminate which extant species they affect (as only a tiny fraction of species is currently in the genomic/orthology databases). We then propose to test whether gene duplications correlate with species diversification. In addition, our resulting duplication rate estimation could also be correlated with phenotypic traits, using appropriate phylogenetic comparative methods.

Keywords: gene duplication, phylogeny, vertebrates

*Speaker

†Corresponding author: guillaume.louvel@ens.fr

IDENTIFICATION OF CORE-GENOME OF ASIAN RICE AND THE IMPACTS OF DOMESTICATION ON THIS CORE-GENOME

Nguyet Dang *† 1, Francois Sabot 2

1 University of Science and Technology of Hanoi (USTH) – 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam

2 Institut de Recherche pour le Développement (IRD [France-Sud]) – Institut de recherche pour le développement [IRD] – 911 avenue Agropolis,BP 6450134394 Montpellier cedex 5, France

Background: With the improvement of sequencing technology, the establishment of reference genomes has been accelerated, hence, has contributed to various studies from functional genomics to genetic variations. However, using only one reference genome to represent a species is not sufficient. Comparing four random genomic regions between maize inbred lines B73 and Mo17 showed that only 50% of the sequences are shared (Brunner, 2005). Therefore, in certain situations, it is necessary to shift from reference genome to pan-genome, which is a representation of all available genomic content in a certain species or in a studied population. A pan-genome consists of three compartments: a core-genome shared among all individuals, dispensable-genomes composing common genes in some but not all individuals and individual-specific-genomes (Tettelin et al., 2005). There have been many pan-genomics studies performed on microorganisms (Vernikos et al., 2015). In plants, this approach has been recently carried on wild and cultivated African rice species to investigate the effects of domestication on pan-genome structure (Monat et al., 2018). The publication indicated that there were reductions in the size of the core-genome as well as the dispensable-genomes of cultivated species in comparison with the wild one.

Research objectives: On account of the available data for Asian rice, we would like to apply the same approach to identify pan-genome compartments of cultivated species *Oryzasativa* and its wild counterpart *Oryza rufipogon.* Afterwards, by comparing the two pan-genomes, we would like to determine how domestication affects the evolution of pan-genome in Asian rice.

Methods: The sequences of *Oryza sativa* is obtained from 3K Genome Project (Li et al., 2014) while the data of *Oryza rufipogon* is provided by IRIGIN (http://irigin.org). Then, these sequences are mapped with the reference genome *Oryza Nipponbare* (Kawahara et al., 2013) by using TOGGLe (Tranchant-Dubreuil et al., 2018). Next, the number of reads are counted and normalized, hence, the presence/absence variations are calculated. The sequences are subsequently assembled and annotated. Eventually, each compartment of the pan-genomes is identified and domestication related gene family is classified by GO analysis.

*Speaker

†Corresponding author: thi-minh-nguyet.dang@ird.fr

Expected results: It is expected that our study will firstly give a full overview of the available genetic content in Asian rice. The results coming from pan-genome analysis will be able to help identify the changes in genetic diversity within two species *Oryza sativa* and *Oryza rufipogon*and provide more details about the effects of domestication on rice species, especially Asian rice. Overall, this research supports the idea of using bioinformatics tools in order to contribute to the understandings of rice, a staple plant playing important roles in food security.

References:

Brunner, S.(2005). Evolution of DNA Sequence Nonhomologies among Maize Inbreds. PLANT CELL ONLINE **17**: 343–360.

Kawahara, Y. et al.(2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice **6**: 1–10.

Li, J.-Y., Wang, J., and Zeigler, R.S.(2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience **3**: 8.

Monat, C., Tranchand, C., Engelen, S., Labadie, K., Paradis, E., Tando, N., and Sabot, F.(2018). Comparison of two African rice species through a new pan-genomic approach on massive data. bioRxiv.

Tettelin, H. et al.(2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." Proc. Natl. Acad. Sci. **102**: 13950–13955.

Tranchant-Dubreuil, C., Ravel, S., Monat, C., Sarah, G., Diallo, A., Helou, L., Dereeper, A., Tando, N., Orjuela-Bouniol, J., and Sabot, F.(2018). TOGGLe, a flexible framework for easily building complex workflows and performing robust large-scale NGS analyses. bioRxiv: 245480.

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H.(2015). Ten years of pan-genome analyses. Curr. Opin. Microbiol. **23**: 148–154.

Keywords: Pangenome, Comparative genomics, Evolution, Asian rice

Evaluation of different approaches for plasmid detection from genome assemblies

Brice Letcher * [1,2], Aurélien Griffon * † [1], Ghislaine Guigon [1]

[1] Bio-Mérieux [Marcy lÉtoile] – BIOMERIEUX – 376 Chemin de lÓrme, 69280 Marcy-l´toile, France
[2] Université Claude Bernard Lyon 1 (UCBL) – Université de Lyon, Université Lyon 1 – 43, boulevard du 11 novembre 1918, 69622 Villeurbanne cedex, France

Background

Plasmids are mobile genetic elements carried in the cells of various bacteria. They mediate genetic exchanges between bacteria through conjugation, the process of DNA transfer between bacterial cells. Being mostly non-essential for host survival in regular conditions, they persist in bacterial cells by carrying genes providing a selective advantage in specific environments. Plasmids also possess their own replication initiation and control genes which must be tightly coordinated with host replication for persistence.

Plasmids often carry antimicrobial resistance (AMR) and virulence genes rendering antibiotics inefficient and bacteria more dangerous. In this context, plasmid detection is important for monitoring the global extent and spread of AMR genes through bacterial populations. It can also have an impact on outbreak monitoring in a clinical setting.

While plasmids can be isolated in the lab prior to sequencing using methods such as alkaline lysis or TRACA (Transposon-aided capture), these approaches are restricted to small and circular plasmids. PCR-based plasmid typing is also available but requires curated markers. In contrast, whole genome sequencing of isolates or metagenomic samples provides access to all of the genetic information, but *in silico* methods are required for distinguishing between chromosomal and plasmidic sequences.

It is known that plasmid detection is greatly facilitated if the sequencing data consists of long reads (Hunt *et al.* 2015; George *et al.* 2017), as these produce less fragmented assemblies. High-quality PacBio assemblies would indeed resolve the problem in most instances, producing one chromosomal contig and one to several other contigs each corresponding to a plasmid. However, long reads are either expensive or highly error-prone, and the majority of public sequencing data comes from Illumina short reads. The methodologies studied will address the challenge of detecting plasmids within bacterial *de novo* genome assemblies of short reads.

Existing methods for plasmid detection rely on features distinguishing them from the host chromosome. One line of programs, including PlasmidSPAdes (Antipov *et al.* 2016) and Recycler (Rozov *et al.* 2016), identifies cycles in the graph used for *de novo* assembly. Plasmids are usually circular and have a characteristic copy number that may differ from the host chromosome.

*Speaker

†Corresponding author: aurelien.griffon@biomerieux.com

These methods thus use the read information to find cycles with depth of coverage different from the median coverage of the whole assembly.

Another approach relies on the assumption that sequence composition differs between plasmids and chromosomes. Each contig can be described by a kmer profile, the relative abundance of words of size k in the sequence. Machine learning algorithms can train models to classify contigs as plasmidic or chromosomal based on kmer features. cBar (Zhou and Xu 2010) and PlasFlow (Krawczyk *et al.* 2018) are two publicly available software that implement this.

Another tool, PlasmidFinder (Carattoli *et al.* 2014) uses curated plasmid-specific markers which are detected in sequences using an alignment-based approach. Finally, tools such as PlasmidSeeker (Roosaare *et al.* 2018) and PlasmidTron (Page *et al.* 2017) rely on identifying plasmid-specific sequence by masking certain kmers from the dataset of interest, leaving only putative plasmid kmers for analysis. The masked kmers can be identified using a provided chromosomal reference genome (PlasmidSeeker), or a provided set of control datasets (PlasmidTron).

Here, we assess existing methods and develop different methods for plasmid detection from bacterial WGS assembly. Raw data is assumed not to be systematically available, in order for the method to be applicable if only an assembly is provided as input. We first present sequence-based metrics for assessing differences between chromosomal and plasmidic contigs. Then, we benchmark original approaches against the main existing approaches for plasmid detection.

Methods and Results

First, we looked for ways of classifying a set of contigs with no prior information, which effectively excludes machine learning, plasmid-specific marker databases, and plasmid sequence databases. Assembly graph analysis software was also excluded because read information is required and for our applications, the read information may be missing.

We thus first explored sequence-only based summary statistics in order to discriminate plasmid and chromosome contigs. For this purpose, we built a set of 12 *E.Coli* and 13 *S.aureus* assemblies from public and internal sequencing projects containing 1 to 6 plasmids. Some key features were identified that differed between chromosome and plasmid contigs. Plasmid contigs display a larger mean Euclidean distance to all other contigs on the basis of their kmer profile. Plasmid contigs also have lower relative entropy-a metric derived from information theory- than chromosome contigs, indicating that they were found to be more unpredictable, *ie* less structured. Finally, plasmid contigs seem to display repeat motif enrichment relative to chromosome contigs.

Being only sequence summary statistics, these features were not sufficient for plasmid/chromosome classification using *ad hoc* thresholds or machine learning approaches. They could however be reasonably integrated into machine learning models using both these features and the full kmer profile. Biologically, they also provide some insight as to consistent plasmid/chromosome differences.

Second, we seek to develop alternative methods to the existing software for plasmid detection in order to improve prediction accuracy. We explore using Support Vector Machines (SVM) for classification, in contrast to the use of neural networks in the PlasFlow software (Krawczyk *et al.* 2018). We also implement an in-house tool for plasmid detection using the PlasmidFinder database with post-processing optimization. Finally, we evaluate alignment–based approaches to match contigs to a large chromosome and plasmid sequence database using Blast followed by specific post-processing based on sequence length.

To assess our alternatives, we use a dataset recently published for benchmarking existing plasmid detection methods (Arredondo-Alonso *et al.* 2017). The published benchmark compares plasmidSPAdes, Recycler, cBar and PlasmidFinder. The authors of PlasFlow also used this dataset to assess its performance, yielding a clear benchmark of most existing methods. To compare our approaches with current existing tools fairly we build a similar sequence database as the one built by the authors of PlasFlow for model training, with accessions in the test dataset removed from the train database.

On our SVM and our Blast-alignment approaches, we obtain superior performance compared to the best existing methods. The best reported performances are 86% for plasmidic recall (fraction of reference plasmid recovered) obtained by PlasFlow, and 75% plasmidic precision (1-fraction of chromosome sequence assigned plasmidic) obtained by plasmidSPAdes. Using our SVM approach we obtain 87% recall and 75.6% precision, and with the Blast-alignment approach we obtain 89.8% recall and 79.8% precision. Note that PlasmidFinder achieves 92% precision but with a recall of only 36%, because it is based on a limited number of plasmid-specific markers.

Conclusion

In this work, we highlight sequence-based summary statistics which broadly differ between chromosomes and plasmids. These could be integrated as features in more complex machine learning models and they provide a basis for future investigation. How consistent these features are across different bacterial taxa, different numbers of plasmids per cell, or different plasmid families is not known and would be to pursue.

We also study and extend an existing benchmark of the main methods for plasmid detection from an assembled set of contigs and provide alternative performant methods. Machine Learning using SVM and Blast alignments followed by post-processing of results perform on a superior level to state of the art tools. Additional benefits include that our linear SVM is a simpler model than PlasFlow's neural network and that we do not require read information unlike plasmidSPAdes. The main advantage of sequence databases is an increase in performance with the addition of more sequences, provided that reasonable time and memory scalability are possible.

In the context of an increasing use of next-generation sequencing for clinical microbial diagnostics, developing appropriate *in silico* methods for the detection of plasmids is required. These can help face the increasing prevalence of antibiotic resistance worldwide and an increasing need for monitoring and controlling outbreaks.
In terms of future improvement, methods relying on extensive plasmid sequence databases, including machine learning, plasmid marker detection, and plasmid database searching, can gain from the ongoing increase in publicly available curated plasmid sequences. We find that plasmid reconstruction from short-read sequencing data is imperfect, highlighting the potential benefits of long-read data. All existing methods assigning a set of contigs as plasmidic or chromosomal would naturally extend to long reads with little to no modification required.

Keywords: plasmid, plasmids, plasmide, plasmides, detection, methods, NGS, short reads, tools, bioinformatics, Blast, Machine learning, information theory, de novo assembly

ProkTE: A web server for large-scale analysis of genetic context associated to the insertion sequences and their putative roles

Sebastien Tempel *† [1], Justin Bedo [2], Mike Chandler [3], Emmanuel Talla * [4]

[1] Laboratoire de Chimie Bacterienne (LCB UMR7283) – CNRS : UMR7283, Aix-Marseille Université - AMU – 31 Chemin Joseph Aiguier 13009 Marseille - France, France
[2] Walter and Eliza Hall Institute (WEHI) – 1G Royal Parade, Parkville VIC 3052, Australia
[3] Laboratoire de microbiologie et génétique moléculaires (LMGM) – Université Paul Sabatier - Toulouse 3, Centre National de la Recherche Scientifique : UMR5100 – 118 route de Narbonne 31062 Toulouse cedex 9, France
[4] Laboratoire de Chimie Bacterienne (LCB UMR7283) – CNRS : UMR7283, Aix-Marseille Université - AMU – 31 Chemin Joseph Aiguier 13009 Marseille, France

Insertion sequences (IS) are small and simple transposable elements that are widely disseminated in microbial genomes [1]. IS are mainly responsible for the mutations and recombinations in prokaryotic genomes and also participate to the gene regulatory networks as promoters or transcription factor binding sites (TFBS) [2, 3]. From 4729 known IS (classified in 30 superfamilies) located in ISFinder webserver (www-is.biotoul.fr) [1] and 5196 annotated organisms from NCBI database, we identified of all IS occurrences in prokaryotic genomes. These results were combined to regulatory sequences from various TFBS databases and lead to a website (ProkTE, http://lcb.cnrs-mrs.fr/ProkTE/index.cgi) that presents the impact and putative functional roles of IS in prokaryotic organisms. This website allows the user (*i*) to dynamically explore the IS occurrences with its genetic environment, (*ii*) to display all synthenic IS-Gene couples, and (*iii*) predict the IS role on the neighboring genes.

Siguier P, et al. 2006. Nucleic Acids Res. 34 :D32-6.

Dziewit L, et al. 2012. PLoS One. 7(2) :e32277.

Uliczka F, et al. 2011. PLoS Pathog. 7(7) :e1002117.

Keywords: Insertion sequence : regulatory sequence : web server

*Speaker

†Corresponding author: sebastien.tempel@univ-amu.fr

A high-quality genomic database for clinical pathogens outbreak monitoring using NGS technology

Aurélien Griffon *† [1], Gaël Kaneko , Caroline Mirande , Margaux Chapel , Bruno Muller , Marilyne Rumigny Pierrot , Emmanuelle Santiago Allexant , Ghislaine Guigon

[1] Bio-Mérieux [Marcy lÉtoile] (BIOMERIEUX) – BIOMERIEUX – 376 Chemin de l'Orme, 69280 Marcy-l'Étoile, France

Background: Surveillance is essential to prevent Healthcare-Associated Infections (HAI) and the spread of pathogens in healthcare institutions. Medical records and strain typing together provide fundamental clues to detect HAI-outbreaks as soon as possible. For several years, PFGE (Pulsed-Field Gel Electrophoresis) and MLST (MultiLocus Sequence Typing) have been the widely used typing methods associated to resistance characterization for outbreaks investigation. However, due to the limited resolution of these traditional methods and their labor intensive aspect, they are difficult to standardize and some outbreaks may be difficult to confirm. The drastic reduction in the cost of Next-Generation Sequencing (NGS) technology and the possibility to sequence the whole genome of bacteria in routine contribute to solve this issue by providing the ultimate resolution for strain typing at an increasingly low cost.
To be effective, this novel whole-genome approach requires the development of a high quality epidemiologically-oriented database of genomes (EpiKB). It should cover the genetic diversity of circulating clones in order to explore the (dis)similarities between suspected and unrelated strains.

To build this EpiKB, the contribution of public databases is essential. Some of them proposed well-annotated reference genomes such as PATRIC or NCBI Genbank or RefSeq, whereas others are specifically designed to store raw sequencing data. For several years, the last ones surpassed the first ones in terms of data quantity, following the popularization of NGS technology, but not in terms of quality. Therefore, exploring and filtering public data is fundamental before constructing any reference database. Moreover, despite the very large amount of data available in public databases, big biases remain regarding the genomic diversity of organisms. We thus also recruited specific clones involved in HAI in order to improve the epidemiologically-oriented database.

Methods: Leveraging our expertise coming from the first BIOMÉRIEUX EPISEQ® application, focused on the surveillance of infections due to *Staphylococcus aureus*, we extended the database to 12 other most prevalent HAI-related organisms: *Acinetobacter baumannii*, *Burkholderia cepacia complex*, *Clostridioides* (ex-*Clostridium*) *difficile*, *Klebsiella* (ex-*Enterobacter*)

*Speaker

†Corresponding author: aurelien.griffon@biomerieux.com

aerogenes, *Enterobacter cloacae complex*, *Enterococcus faecalis*, *Enterococcus faecium*, *Escherichia coli & Shigella spp. complex*, *Klebsiella oxytoca*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa* and *Serratia marcescens.*

In order to build the reference database over all species, 8802 genomes were downloaded from PATRIC and/or NCBI RefSeq databases. They mainly correspond to complete genomes, non-epidemiologically related. We also retrieved 53946 datasets coming from the Sequence Read Archive (NCBI/SRA) together with their epidemiological information extracted from the European Nucleotide Archive (EBI/ENA) to enrich the epidemiological database and better cover the genomic diversity of species. About 50% of these public NGS datasets were filtered out based on sequencing platforms and models, theoretical coverage depth and read length to avoid poor quality data. Finally, 1169 well-characterized strains related to outbreaks coming from National Reference Centers and hospitals were collected with their epidemiological metadata to extend the database. They correspond to specific strains having a Sequence Type (ST) prevalent in HAI (Mirande et al. 2017). In order to standardize the database, only Illumina NGS datasets were retrieved from public databases and strains collected were sequenced using Illumina HiSeq or MiSeq sequencers. The initial database contained 36923 genomes coming from the three sources: 1) public genomes, 2) public raw sequencing data analysis and 3) sequencing and analysis of collected strains.

The metadata retrieved includes, in particular, the collection date, the collection location, the host of the pathogen, the disease associated and above all, the epidemiological link between strains. Several steps of collection, filtering, standardization and curation of the metadata were carried out to obtain an exhaustive and a high-quality epidemiological database. All epidemiological links between strains were validated using both a whole-genome MLST (wgMLST) distance tree and the metadata.

Results: An automated and standardized process was developed to analyze collected and public strains, including preprocessing of reads, *de novo* genome assembly and quality assessment. From the initial database, about 14% of genomes were discarded in the final reference database after performing the quality assessment including reads and assembly statistics evaluation, species validation, contamination detection and duplicates removing.

The final database contains 31862 genomes over all species (from > 100 for rare species up to > 6000 for common species) with their curated epidemiological information. About 8% of these genomes correspond to well-characterized outbreaks over which collected strains represent 30%. The high number of genomes, including non-related strains, brings a large biological diversity to the database, thus covering 80 prevalent sequence types of pathogens involved in HAI worldwide (i.e. 83% of prevalent STs over the 13 species). The geographical diversity supported by the database includes Europe (23% of genomes with known origin), North America (15%), Asia (2%) and in a lesser extent Africa and South America. The validation of all genomes, epidemiological information and epidemiological links between strains ensures the quality of the database in order to provide relevant and efficient results for a clinical usage.
Conclusion: In order to take advantage of the ultimate resolution for strain typing enabled by NGS technology, we developed the first database of genomes involving the 13 most prevalent HAI-related species. This unique, curated and truly extensive collection of 31862 reference genomes together with their epidemiological information brings an essential component to explore (dis)similarities between strains for clinical pathogens outbreak monitoring. The database covers a wide biological and geographical diversity, includes confirmed and well-characterized outbreaks, thus facilitating the interpretation of NGS-based typing results associated to the BIOMÉRIEUX EPISEQ® CS application. A continuous effort is made to maintain, update and improve this epidemiologically-oriented database, particularly for rare HAI-related species.

Keywords: NGS, sequencing, database, outbreak monitoring, epidemiology, HAI, healthcare associated infections

PathostDB - Base de donnée légère d'interactions hôtes-pathogènes de plantes

Maira Barca [*†] , François Sabot[‡ 1]

[1] Diversité, adaptation, développement des plantes (DIADE) – Institut de recherche pour le développement [IRD] : UR232, Université Montpellier II - Sciences et techniques – Centre IRD de Montpellier 911 av Agropolis BP 604501 34394 Montpellier cedex 5, France

PathostDB est une base de données légère d'interactions hôtes-pathogènes de plantes, *via* une interface de requêtage à partir de données sous fichiers plats. Elle permet d'accéder à des informations de résistance et/ou de sensibilité pour divers pathogènes. Les résultats obtenus sont visibles sous forme de tableaux sur une interface web.
Ce projet est né suite à l'obtention de nombreuses données de résistance et/ou sensibilité de divers pathogènes sur une collection de riz Africain dans le cadre du projet MENERGEP.

Le projet MENERGEP (*"Methodologies and new resources for genotyping and phenotyping of African rice species and their pathogens for developing strategic disease resistance breeding programs"*) a été porté par Africa Rice, l'IRD et le CIRAD et financé par le GRiSP de 2012 à 2014. Ce projet avait pour but la caractérisation de divers pathogènes du riz en Afrique et permettre l'identification de sources de résistance dans la diversité naturelle du riz. Pour ce faire, une collection d'accessions de riz Africain a été utilisée comprenant des accessions de riz sauvages (*Oryzae barthii*) et de riz cultivés (*Oryzae glaberrima*).

Les données ont été obtenues sur de nombreux sites géographiques dans divers pays d'Afrique, notamment au Bénin, Burkina Faso, Cameroun, Sénégal.

Afin de permettre la visualisation de cet ensemble de données, une base de données d'interactions spécifique au riz Africain a vu le jour : la base MENERGEPdb. Cette dernière permettait ainsi un recensement des diverses interactions entre plusieurs pathotypes de la bactérie Xanthomonas et du virus RYMV face aux accessions de riz sauvages et cultivées.

Ces pathogènes peuvent avoir une portée dévastatrice sur le rendement (destruction de récoltes) et la sécurité alimentaire de certains pays du Sud. C'est pour cette raison que l'identification et l'accessibilité des facteurs de résistances à ces pathogènes sont importantes.

La base de données MENERGEPdb a ensuite évolué pour devenir PathostDB, qui se veut être une base de données générique et légère pouvant être utilisée pour tous types d'interactions hôtes-pathogènes à la différence de la base MENERGEPdb étant spécifique aux données de riz Africain. PathostDB pourra ainsi être déployé sur divers sites indépendamment des données utilisées.

[*] Speaker

[†] Corresponding author: mairaxb@yahoo.fr

[‡] Corresponding author: francois.sabot@ird.fr

Les bases de données à partir de fichiers plats, ou *flat database*, sont des bases de données composées de fichiers textes ou CSV. Les fichiers plats contiennent généralement un enregistrement par ligne et peuvent être délimités en différents champs par des caractères spéciaux. C'est le cas notamment pour le format CSV où les champs sont séparés par des virgules.

Le choix de l'architecture de la base s'est portée sur celle à partir de fichiers plats. En effet, pour ce projet on recherche un système portable avec peu de maintenance nécessaire et ne nécessitant pas de personnel formé pour la gestion de la base. Les base de données relationnelles ou NoSQL possèdent de nombreux avantages, notamment au niveau des systèmes de requêtage étant plus complets et de sécurité des données, cependant elles nécessitent une maintenance et un personnel formé en fonction du type de la base. En revanche, les bases de données à partir de fichiers plats ont les caractéristiques recherchées, bien qu'un système de requête soit plus difficile à mettre en place.

La base PathostDB va être crée à partir de trois fichiers CSV : un fichier contenant des informations sur les plantes qui seront les hôtes, un fichier contenant des informations sur les pathogènes et les pathothypes qui leurs sont associés et un dernier fichier contenant les informations de résistance et/ou sensibilité de ces pathogènes. Ces fichiers utilisés comme fichiers d'entrées vont ensuite être transformés en fichiers JSON et *hash*, afin d'être plus manipulable pour l'implémentation du système de requêtes.

Le langage Perl-CGI est choisi car il permet une implémentation à la fois de la partie serveur mais également de l'interface web.

La base de donnée doit être stockée sur un serveur type Apache afin d'être fonctionnelle. Sur ce serveur, seront stockés les fichiers plats de la base, l'ensemble des scripts pour la création et manipulation des données, ainsi que les fichiers de style et fichiers temporaires.

L'initialisation et l'enrichissement de la base sont gérés par un script prenant en entrée le répertoire contenant les fichiers d'entrées. Lors d'une première utilisation, il y aura création des fichiers JSON et *hash* associés, permettant ainsi une manipulation des données plus facilement. Si l'on souhaite enrichir la base par la suite, il suffit simplement de rajouter des fichiers dans ce répertoire et les informations seront alors rajoutés aux fichiers JSON et hash précédemment crées.

Une des importantes fonctionnalités de cet outil est la restriction des données. En effet, tous les utilisateurs ne peuvent pas avoir accès à l'ensemble des informations de la base de données. Certains utilisateurs auront la possibilité de se connecter au système avec un jeu de *login* et mot de passe, et auront ainsi accès à l'ensemble des informations. De plus seuls les utilisateurs connectés à la base pourront déposer des fichiers *templates* d'enrichissement de la base.

L'interface web est écrit en Perl-CGI et est composée de plusieurs onglets notamment des onglets donnant des informations sur les plantes ainsi que les pathogènes, mais aussi des onglets de requêtes. Deux onglets de requêtes sont présents : un donnant des informations basiques, et un autre donnant plus d'informations notamment sur les pathogènes.

Les onglets de requêtes de l'interface web vont être dynamiques grâce aux scripts javascript et ajax associés. En effet, après lancement d'une requête (recherche de toutes les variétés de *O. barthii* résistantes à RYMV par exemple), le script gérant les ajax reçoit les paramètres des éléments à modifier ou afficher dans la page HTML de départ par l'intermédiaire d'une fonction javascript. Ceci permet l'affichage des résultats des différentes requêtes sur une partie de la page

web sans avoir à recharger la page entière.

PathostDB permet également de donner des différentes informations sur les plantes et pathogènes, telles que la localisation et l'abondance d'un pathogène en fonction de la variété de plante, des informations statistiques (proportion de résistance/sensibilité pour chaque pathogènes), des images des pathogènes et/ou des symptômes de ces derniers sur les plantes ainsi que la localisation des différents pathogènes en fonction de leurs coordonnées géographiques exactes si précisées ou en fonction du pays du pathogène en question dans le cas contraire.
L'outil PathostDB est donc une interface simple d'utilisation, utilisable sans pré-requis de formation au management de base de données et permettant rapidement l'accès à des informations de résistance.

Keywords: database, plants, host, pathogens, resistance

Using metabarcoding data to reveal the distribution and interactions of cyanobacteria in the oceans

Ewen Corre * [1,2], Laura Rubinat [1], Nicolas Henry [1], Eric Pelletier [2], Colomban De Vargas [1]

[1] Adaptation et diversité en milieu marin (AD2M) – Sorbonne Université, UPMC, CNRS : UMR7144 – Station Biologique de Roscoff - Place Georges Teissier - BP 74 29682 ROSCOFF CEDEX, France
[2] Genomics Metabolics (GM) – Commissariat à l'énergie atomique et aux énergies alternatives : DSV/IG, CNRS : UMR8030 – Genoscope - 2 rue Gaston Crémieux CP5706 91057 EVRY Cedex, France

Cyanobacteria form a large group of photosynthetic prokaryotes that can be characterized by their high genetic diversity and their broad range of habitat across latitudes. A significant part of them can also be found as endosymbionts within other organisms, such as lichens, plants and unicellular eukaryotes (protists).

In this study, we aimed at revealing the diversity and ecological distribution (including symbiotic interactions) of marine cyanobacteria across the oceans.

We analyzed 16S ribosomal DNA sequences (*i.e.* prokaryotic and chloroplastic metabarcodes) from more than a thousand size-fractionated plankton communites collected during the *Tara* Oceans expedition. First, using a combination of phylogenetic placement and ecological analyses, we assessed the diversity and the distribution of marine cyanobacteria. Further, we integrated ecological and genetic data to identify groups potentially involved in symbiotic relationships with other organisms.

We found that despite the diversity of cyanobacteria, almost all families have marine representatives. Some picocyanobacteria (_˜ 1 μm large) were also found in the micro- and mesoplankton ($>$ 20 μm in size), suggesting that they were associated with larger organisms. These last predictions can and will be compared with the *Tara* Oceans imaging datasets to be validated or invalidated.

Keywords: metabarcoding, ecology, cyanobacteria, phylogenetic placement, symbiosis, oceans

*Speaker

High intraspecific genome diversity in the model arbuscular mycorrhizal symbiont Rhizophagus irregularis

Emmanuelle Morin * [1]

[1] INRA Grand Est (UMR 1136 IAM) – Institut National de la Recherche Agronomique – route d'Amance 54280 Champenoux, France

Arbuscular mycorrhizal fungi (AMF) are known to improve plant fitness through the establishment of mycorrhizal symbioses. Genetic and phenotypic variations among closely related AMF isolates can significantly affect plant growth, but the genomic changes underlying this variability are unclear. To address this issue, we improved the genome assembly and gene annotation of the model strain Rhizophagus irregularis DAOM197198, and compared its gene content with five isolates of R. irregularis sampled in the same field. All isolates harbor striking genome variations, with large numbers of isolate-specific genes, gene family expansions, and evidence of interisolate genetic exchange. The observed variability affects all gene ontology terms and PFAM protein domains, as well as putative mycorrhiza-induced small secreted effector-like proteins and other symbiosis differentially expressed genes. High variability is also found in active transposable elements. Overall, these findings indicate a substantial divergence in the functioning capacity of isolates harvested from the same field, and thus their genetic potential for adaptation to biotic and abiotic changes. Our data also provide a first glimpse into the genome diversity that resides within natural populations of these symbionts, and open avenues for future analyses of plant-AMF interactions that link AMF genome variation with plant phenotype and fitness.

Keywords: comparative genomics, intraspecific variation, pan, genome, genome annotation

*Speaker

AquaPony: visualization of phylogeographic information on phylogenies

Bastien Cazaux* [1,2], Guillaume Castel [3,4], Eric Rivals † [5,6]

[1] Kapteyn Astronomical Institute, University of Groningen – P.O. box 800 9700AV Groningen, Netherlands
[2] Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier : UMR5506, Centre National de la Recherche Scientifique : UMR5506 – 161 rue Ada - 34095 Montpellier, France
[3] Institut d'Electronique et de Télécommunications de Rennes (IETR) – CNRS : UMR6164, Université de Rennes I, Institut National des Sciences Appliquées de Rennes, SUPELEC – Campus de Beaulieu Bâtiment 11D 35042 Rennes Cedex, France
[4] Centre de Biologie pour la Gestion des Populations (CBGP) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement : UMR55, Centre international d´tudes supérieures en sciences agronomiques : UMR1062, Institut national de la recherche agronomique [Montpellier] : UMR1062, Université de Montpellier : UMR1062, Institut de Recherche pour le Développement : UMR1062, Institut national d'études supérieures agronomiques de Montpellier, Centre international d´tudes supérieures en sciences agronomiques : UMR1062, Centre international d´tudes supérieures en sciences agronomiques : UMR1062, Centre international d´tudes supérieures en sciences agronomiques : UMR1062, Centre international d'études supérieures en sciences agronomiques : UMR1062 – 755 avenue du Campus Agropolis, 34988 Montferrier sur Lez, France
[5] Institut de Biologie Computationnelle (IBC) – Univrsité de Montpellier – Montpellier, France
[6] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) – CNRS : UMR5506, Université Montpellier II - Sciences et Techniques du Languedoc – Montpellier, France

The visualization and interpretation of evolutionary spatiotemporal scenarios is broadly and increasingly used in infectious disease research, ecology, or agronomy. Using probabilistic frameworks, well-known tools can infer from molecular data ancestral traits for internal nodes in a phylogeny, and numerous phylogenetic rendering tools can display such evolutionary trees. However, visualizing such ancestral information and its uncertainty on the tree remains tedious. For instance, ancestral nodes can be associated to several geographical annotations with close probabilities and thus, several migration or transmission scenarios exist. We expose a web-based tool, named AquaPony, that facilitates such operations. Given an evolutionary tree with geographic annotations, the user can easily control the display of ancestral information on the entire tree or a subtree, and can view alternative phylogeographic scenarios along a branch according to a chosen uncertainty threshold. AquaPony interactively visualizes the tree and eases the objective interpretation of evolutionary scenarios. AquaPony's implementation makes it highly responsive to user interaction, which instantaneously update the tree visualizations even for large trees (which can be exported as image files).

*Corresponding author: cazaux@cs.helsinki.fi

†Speaker

Keywords: phylogeography, inference, uncertainty, tree, web, software, interactive visualisation

Développement d'outils pour la détection de variants génomiques dans le cadre d'analyses de pan-génomes

Amad Diouf * [1,2], François Sabot [3,4], Christine Tranchant-Dubreuil [3,5]

[1] Master Sciences et Numérique pour la Santé, Bioinformatique Connaissances Données, Année M2, Université de Montpellier, (M2 BCD) – Master 2 Bioformatique, Connaissances, Données – Campus Triolet, Place Eugène Bataillon, 34095, Montpellier Cedex 5, France, France

[2] Diversité, adaptation, développement des plantes (DIADE) – Université de Montpellier, Institut de Recherche pour le Développement – Centre IRD de Montpellier 911 av Agropolis BP 604501 34394 Montpellier cedex 5, France

[3] Diversité, adaptation, développement des plantes (DIADE) – Institut de recherche pour le développement [IRD] : UR232, Université Montpellier II - Sciences et techniques – Centre IRD de Montpellier 911 av Agropolis BP 604501 34394 Montpellier cedex 5, France

[4] South Green (SG) – Institut de recherche pour le développement [IRD], Institut national de la recherche agronomique (INRA), CIRAD, Bioversity – France

[5] South Green (SG) – Institut de recherche pour le développement [IRD], Institut national de la recherche agronomique (INRA), CIRAD, Bioversity – France

Les variations structurales sont de différents types et peuvent être très complexes. Dans le cadre d'études de diversité basées sur ces variations, un des premiers obstacles est leur détection. De nombreux outils de détection des variants, basés sur des données de séquençage haut débit, ont été mis à disposition de la communauté scientifique Ces outils sont souvent spécifiques à un type de variation et peuvent rencontrer des limites que cela soit au niveau de leur spécificité ou de leur sensibilité [1].

Une intégration de plusieurs outils a été ici mise en place, afin de réaliser une détection à grande spécificité et sensibilité des variations structurales. Une première étape a été de recenser un grand nombre d'outils disponibles puis une seconde étape de présélection a été faite pour retenir les outils les plus utilisés par la communauté. Une phase de tests a permis par la suite de ne retenir que les outils aux meilleures spécificités et sensibilités. Ceux-ci ont été intégrés au gestionnaire de workflows de la plateforme South Green, TOGGLe [2]. Les outils intégrés se basent sur l'assemblage des reads d'un échantillon à comparer ou sur l'alignement de ceux-ci (qui permet alors une analyse de la profondeur de couverture ou du profil d'alignement discordant) [3]. Une bioanalyse dans le contexte d'une étude de la diversité a été réalisée avec cette brique logicielle ainsi obtenue afin de montrer la progression qu'elle marque dans la thématique de la détection des variations structurales.

Références:

L. Tattini, R. D'Aurizio, and A. Magi, "Detection of Genomic Structural Variants from Next-Generation Sequencing Data.," Front. Bioeng. Biotechnol., vol. 3, p. 92, 2015.

C. Tranchant-Dubreuil et al., "TOGGLe, a flexible framework for easily building complex workflows and performing robust large-scale NGS analyses," bioRxiv, p. 245480, 2018.

*Speaker

P. Guan and W.-K. Sung, ”Structural variation detection using next-generation sequencing data,” Methods, vol. 102, pp. 36–49, Jun. 2016.

Keywords: variations structurales, technologies de séquençage haut débit, intégration, développement, bio, analyse.

Mise au point d'approches d'analyses génomiques de novo avec la technologie long read Oxford Nanopore

Valentin Klein *† [1], François Sabot [2,3], Sébastien Cunnac [4]

[1] Diversité, adaptation, développement des plantes (DIADE) – Institut de recherche pour le développement [IRD] : UMR232, Université de Montpellier – Centre IRD de Montpellier 911 av Agropolis BP 604501 34394 Montpellier cedex 5, France

[2] Diversité, adaptation, développement des plantes (DIADE) – Institut de recherche pour le développement [IRD] : UMR232, Université de Montpellier – Centre IRD de Montpellier 911 av Agropolis BP 604501 34394 Montpellier cedex 5, France

[3] South Green (SG) – Institut de recherche pour le développement [IRD], Institut national de la recherche agronomique (INRA), CIRAD, Bioversity – France

[4] UMR - Interactions Plantes Microorganismes Environnement (UMR IPME) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Université de Montpellier, Institut de Recherche pour le Développement : UMR186 – IRD France-Sud 911, avenue Agropolis BP 64501 34394 Montpellier cedex 5, France

L'étude de la diversité génomique des espèces vivantes est un outil permettant de mieux cerner la structure des génomes, l'apparition et le fonctionnement des structures codantes et régulatrices et en fin de compte mieux comprendre les mécanismes du vivant et son évolution. Les technologies de séquençage de l'ADN permettent de collecter des grands volumes d'informations génomiques et propulsent la recherche dans ce domaine. Bien que traditionnellement des technologies de séquençage de seconde génération à reads courts soient utilisées, des nouvelles technologies de troisième génération produisant des *reads* plus longs mais moins précis sont disponibles et permettent d'acquérir une meilleure vision d'ensemble des génomes, notamment dans les régions traditionnellement difficiles à séquencer. Ces reads plus longs ouvrent des nouvelles perspectives dans l'assemblage de génomes complet et dans la détection des variation structurales, par exemple.

La technologie de séquençage Oxford Nanopore est une technologie de séquençage *long read* de troisième génération. Elle permet de produire des reads avec une taille moyenne atteignant facilement les 10 à 15kb et une taille maximum dépassant les 1Mb, avec une barrière d'entrée financière et matérielle potentiellement moins importante que pour d'autres technologies long read, et une portabilité unique (le séquenceur MinION fonctionnant sur un port USB classique) qui permet d'envisager des applications sur le terrain. Cela ouvre donc de nouveaux horizons dans l'étude de la diversité génétique des espèces, notamment pour l'amélioration des assemblages *de novo*, la détection des variations structurales à grande échelle, ou encore l'haplotypage. Cette technologie apporte également de nouvelles fonctionnalités uniques comme le RNA-seq direct (sans RT-PCR), ou encore la possibilité d'enrichir une librairie sur une séquence cible en direct sans capture préalable.

*Speaker

†Corresponding author: valentin.klein@etu.umontpellier.fr

L'analyse post-séquençage des données Nanopore demande une révision des *workflows* d'analyse existants. En effet, beaucoup d'outils et d'algorithmes élaborés pour Illumina ne sont pas adaptés au reads longs ou ne tirent pas pleinement partie de la technologie. De nombreux nouveaux outils d'assemblage, de mapping spécifiques aux reads longs ou à Nanopore ont été développés [1], mais en l'absence d'un *gold standard* établi dans un écosystème encore jeune, une évaluation détaillée de ces outils à du être réalisée, suite à quoi des workflows d'analyse ont été mis au point, en appui sur plusieurs projets de recherche.

En effet, sein de l'IRD, plusieurs projets de recherche tirent partie des apports de ces nouvelles technologies pour l'assemblage de génomes de novo et la détection de variations structurales. Des assemblages de bonne qualité des chloroplastes de plusieurs espèces végétales ont été produits, ce qui était impossible à réaliser avec des reads courts en raison de la présence de deux grandes régions inversées répétées fortement conservées.

En outre, une référence de bonne qualité pour une espèce du genre *Paspalum*, diplo´ide et fortement hétérozygote est en cours d'élaboration en utilisant une approche hybride Nanopore/Illumina. Elle servira de socle pour l'étude de mécanismes de reproduction apomictiques dans d'autres espèces du même genre de plo´idie supérieure.

Enfin, un projet de séquençage à grande échelle de génomes de riz africains et asiatiques cherche à mettre en évidence des variations structurales à l'échelle du génome.

Les nouveaux outils sélectionnés et les protocoles d'analyse ont été intégrés dans TOGGLe [2], le gestionnaire de Workflow développé par la plateforme SouthGreen, dont fait partie l'IRD. Ce gestionnaire, orienté vers les utilisateurs non-informaticiens mais qui permet quand même de mettre au point des workflows complexes et de traiter des grandes quantités de données (contrairement aux gestionnaires basés sur des interfaces web), était jusqu'ici concentré sur les analyses Illumina. Cette intégration permet aux chercheurs de modifier facilement les workflows et de lancer leur analyses dans un environnement HPC (High Performance Computing) de manière automatique.

Fritz J Sedlazeck et al. " Piercing the dark matter : bioinformatics of long-range sequencing and mapping ". In : Nature Reviews Genetics (2018). doi : 10.1038/s41576-018-0003-4.

Christine Tranchant-Dubreuil et al. " TOGGLe, a flexible framework for easily building complex workflows and performing robust large-scale NGS analyses ". In : bioRxiv (fév. 2018). doi : 10.1101/245480.

Remerciements : Cette participation à JOBIM, dans le cadre de la seconde année de Master Sciences et Numérique pour la Santé, Parcours Bioinformatique, Connaissances, Données, est financée par le LABEX NUMEV et par l'Institut de Recherche pour le Développement

Keywords: long read, nanopore, assemblage, génomique

A new multiple nucleotide sequence exact search tool for performing fast core-genome Multi-Locus Sequence Typing

Yoann Dufresne *† [1], Valérie Bouchez [2], Alexis Criscuolo [1,3]

[1] Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 25-28 Rue du Docteur Roux, 75015 Paris, France
[2] Unité de Biodiversité et Epidémiologie des Bactéries Pathogènes (BEBP) – Institut Pasteur de Paris – 25-28 Rue du Docteur Roux, 75015 Paris - France, France
[3] Pasteur International Bioresources network (PIBnet), Plateforme de Microbiologie Mutualisée (P2M) – Institut Pasteur de Paris – 25-28 Rue du Docteur Roux, 75015 Paris - France, France

Multi-Locus Sequence Typing (MLST; Maiden et al. 1998) is a standard technique in molecular epidemiology allowing a precise and reproducible genotypic classification of bacterial strains based on several loci (generally, from five to ten housekeeping genes). Most of the common causative agent for bacterial infectious diseases are characterized by a MLST scheme (see e.g. pubmlst.org/databases.shtml). Every scheme is defined by different loci as well as, for each locus, a distinct allele sequence set. In practice, strain genotyping is performed by searching within its genome the allele sequence for each locus of the associated MLST scheme, therefore determining an allelic profile. This simple genotyping approach leads to a standardized and deterministic nomenclature that facilitates communication among microbiologists and contributes to the fast resolution of epidemiological outbreaks.
As sequencing and assembling bacterial genomes are becoming a routine procedure in many labs, MLST approach was recently extended to the set of loci conserved within a considered species in order to better understand its evolution and discriminate closely related strains (Bialek-Davenet et al. 2014, de Been et al. 2015, Moura et al. 2016, Ghanem et al. 2017, Gonzalez-Escalona et al. 2017, Higgins et al. 2017, Bletz et al. 2018, Bouchez et al. 2018). These so-called core-gene MLST (cgMLST) schemes are therefore based on a large number of genes (e.g. from hundreds to thousands of loci, each containing from tens to hundreds of allele sequences). In practice, when dealing with such large-sized cgMLST schemes, running time required to perform allele tagging on a large number of assembled genomes can be important. Indeed, current tools for searching multiple allele sequence sets against genome contigs are only based on BLAST similarity searches (see Page et al. 2017).

We therefore implemented a new bioinformatics tool able to quickly find all occurrences of a large set of nucleotide allele sequences within assembled genome contigs. Based on an adaptation of the Karp and Rabin (1987) algorithm for multiple pattern search, each allele sequence (as well as its reverse-complement) is first broken down into a list of distinct k-mers (i.e. oligonucleotides of length k) to be next characterized by its most representative one. Therefore, an allele match is verified only when its representative k-mer is found within the genome contigs.

*Speaker

†Corresponding author: yoann.dufresne@pasteur.fr

As looking over every k-mer of a nucleotide sequence can be quickly performed via the use of a rolling hash technique, this algorithmic scheme allows very fast running times to be observed.

In order to illustrate the overall performance of our multiple nucleotide sequence exact search tool, we used it for genotyping hundreds of Bordetella pertussis genomes (_˜4.1 Mb each) with the associated cgMLST scheme, made up of 12,536 allele sequences with length varying from 90 to 5,718 nucleotides (Bouchez et al. 2018). We show that our tool accurately achieves allele tagging for these 2,038 loci with faster running time than currently available implementations.

Literature Cited

de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W, Brouwer E, Rogers M, Kraat Y, Bonten M, Corander J, Westh H, Harmsen D, Willems RJ (2015) Core genome multilocus sequence typing scheme for high-resolution typing of Enterococcus faecium. Journal of Clinical Microbiology, doi: 10.1128/JCM.01946-15

Bialek-Davenet S, Criscuolo A, Ailloud F, Passet V, Jones L, Delannoy-Vieillard AS, Garin B, Le Hello S, Arlet G, Nicolas-Chanoine MH, Decré D, Brisse S (2014) Genomic definition of hypervirulent and multidrug-resistant Klebsiella pneumoniae clonal groups. Emerging Infectious Diseases, doi:10.3201/eid2011.140206

Bletz S, Janezic S, Harmsen D, Rupnik M, Mellmann A (2018) Defining and evaluating a core genome multilocus sequence typing scheme for genome-wide typing of Clostridium difficile. Journal of Clinical Microbiology, doi:10.1128/JCM.01987-17

Bouchez V, Guglielmini J, Dazas M, Landier A, Toubiana J, Guillot S, Criscuolo A, Brisse S (2018) Genomic sequencing of Bordetella pertussis for epidemiology and global surveillance of whooping cough. Emerging Infectious Diseases, doi:10.3201/eid2406.171464

Ghanem M, Wang L, Zhang Y, Edwards S, Lu A, Ley D, El-Gazzar M (2017) Core genome multilocus sequence typing: a standardized approach for molecular typing of Mycoplasma gallisepticum. Journal of Clinical Microbiology, doi:10.1128/JCM.01145-17

Gonzalez-Escalona N, Jolley KA, Reed E, Martinez-Urtaza J (2017) Defining a core genome multilocus sequence typing scheme for the global epidemiology of Vibrio parahaemolyticus. Journal of Clinical Microbiology, doi:10.1128/JCM.00227-17

Higgins PG, Prior K, Harmsen D, Seifert H (2017) Development and evaluation of a core genome multilocus typing scheme for whole-genome sequence-based typing of Acinetobacter baumannii. PLoS ONE, doi:10.1371/journal.pone.0179228

Karp RM, Rabin MO (1987) Efficient randomized pattern-matching algorithms. IBM Journal of Research and Development, doi:10.1.1.86.9502

Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenicmicroorganisms. Proceedings of the National Academy of Sciences of USA, doi:10.1073/pnas.95.6.3140

Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, Bj́orkman JT, Dallman T, Reimer A, Enouf V, Larsonneur E, Carleton H, Bracq-Dieye H, Katz LS, Jones L, Touchon M, Tourdjman M, Walker M, Stroika S, Cantinelli T, Chenal-Francisque V, Kucerova Z, Rocha EPC, Nadon C, Grant K, Nielsen EM, Pot B, Gerner-Smidt P, Lecuit M, Brisse S (2016) Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. Nature Microbiology, doi:10.1038/nmicrobiol.2016.185

Page AJ, Alikhan N-F, Carleton HA, Seemann T, Keane JA, Katz LS (2017) Comparison of classical multi-locus sequence typing software for next-generation sequencing data. Microbial Genomics, doi:10.1099/mgen.0.000124

Keywords: Multi Locus Sequence Typing, Bacteria, Software

Comparative study of RNA-Seq analyses based on the genome or transcriptome as reference

Florent Tessier * [1], Olivier Croce

[1] Institut de Recherche sur le Cancer et le Vieillissement (IRCAN) – Université Nice Sophia Antipolis, Institut National de la Santé et de la Recherche Médicale : U1081, Centre National de la Recherche Scientifique : UMR7284 – Faculté de médecine, 28 avenue de Valombrose 06107 Nice Cedex 2, France

RNA Seq is now widely used to identify genes differentially expressed between several biological conditions. Many studies have shown that the results of RNA-Seq analyses can be very different depending on the sequencing technologies or depending on the methods and tools used to carry out these analyses. An important step during the analysis is to choose the reference on which the reads will be aligned to identify the expressed transcripts. This reference is usually the annotated genome of the organism studied, or more rarely its transcriptome when it is available. We studied the differences in results depending on the choice of this reference. We established and compared the list of genes differentially expressed by choosing either the genome or the transcriptome as reference, using nine sets of sequencing data coming from 3 different organisms (human, mouse and zebrafish). We have also extended these comparisons by adding the expression of alternative transcripts for each case.

Our results show that the lists of differentially expressed genes or transcripts are relatively close in both cases. However, it appears that the levels of expression of certain genes or transcripts can sometimes be quite different depending on the reference chosen. Moreover, some rare genes/transcripts are totally absent in one case or another. Therefore, to obtain exhaustive results, we suggest to systematically conduct the 2 types of analyses by considering both the genome and the transcriptome.

Keywords: RNA, seq, transcriptome, genome, DEG analysis

*Speaker

How to improve genome assembly using repetitive elements.

Quentin Delorme * [1], Annie Chateau† [2,3], Anna-Sophie Fiston-Lavier [4], Yasmine Mansour [5,6,7]

[1] Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier : UMR5506, Centre National de la Recherche Scientifique : UMR5506 – 161 rue Ada - 34095 Montpellier, France
[2] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université Montpellier II - Sciences et techniques, CNRS : UMR5506 – CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France
[3] Institut de Biologie Computationnelle (IBC) – CNRS : UMR5506, Université Montpellier II - Sciences et techniques – 95 rue de la Galéra, 34095 Montpellier, France
[4] Institut des Sciences de l'Evolution - Montpellier (ISEM) – CNRS : UMR5554, Institut de recherche pour le développement [IRD] : UMR226 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France
[5] Institut des Sciences de lÉvolution [Montpellier] (ISEM) – Université de Montpellier, Institut de recherche pour le développement [IRD] : UR226, Centre National de la Recherche Scientifique : UMR5554 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France
[6] Institut de Biologie Computationnelle (IBC) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut National de la Recherche Agronomique, Institut National de Recherche en Informatique et en Automatique, Université de Montpellier, Centre National de la Recherche Scientifique – 95 rue de la Galéra, 34095 Montpellier, France
[7] Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier : UMR5506, Centre National de la Recherche Scientifique : UMR5506 – CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France

Repetitive DNA sequences are abundant in almost all species: RRs (Repetitive Regions) may represent up to 90% of genome size [1]. Despite being a fundamental source of genomic diversity and novelty, RRs are responsible of assembly errors yielding bad quality of genome assemblies [2]. Even with advanced high-throughput sequencing technologies, genome assembly is facing a big challenge towards achieving its optimum quality. While reads assembly overcome this issue, often by collapsing or excluding repeats from contigs, scaffolding step ought to handle RRs.

The perspective of this work is to detect, classify and use misassemblies due to RRs to improve genome assemblies. Our hypothesis is that some RRs like Transposable Elements (TEs) are more disruptive elements in the face of genome

assembly process than others, due to their biology. We intend to test whether the assembly errors are more likely caused by long and young TE insertions [3]. We are currently working on Anopheles gambiae's reference genome. Anopheles gambiae is the principal vector of malaria, a disease that afflicts more than 500 million people and causes more than 1 million deaths each year. Improving assemblies may lead to a better understanding of his genome's dynamic and

*Speaker

†Corresponding author: annie.chateau@lirmm.fr

appearance of insecticide resistance. We intend to exploit sequence similarities between repeats family on a three-step process :

- A first step consists to investigate how information on TEs obtained independently of the assembly, could limit their disruptive effects. Using CENSOR [4], we are able to detect different types of RRs dans tag them on contigs.

- In a second step, we put together contigs clusters based on labeled RRs families. This step is meant to reduce possibilities of misjunction between contigs holding two different kind of RRs.

- In each cluster each combinaison of two contigs, leading to the formation of hypothetic scaffolds, is querying against the repeat database Repbase. Thus, scaffolds can be validate by matching with an existing repeat region, leading to the reconstruction of the original sequence.

The aim is to generate scaffold graph from those RRs informations. This graph could be different than scaffold graph based on paired-end reads informations. Here, the challenge will be to confront orientation informations from both graph and try to resolve hypothetic conflict. Algorithmic approach will be developped for evaluation of information relevance.

C. Biemont. A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics, 186(4):1085{1093, Dec 2010.

H.Tang. Genome assembly, rearrangement, and repeats. Chemical Reviews, 107:3391{3406, 2007.

Rajiv C. McCoy, Ryan W. Taylor, Timothy A. Blauwkamp, Joanna L. Kelley, Michael Kertesz, Dmitry Pushkarev, Dmitri A. Petrov, and Anna-Sophie Fiston-Lavier. Illumina truseq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLOS ONE, 9(9):1{13, 09 2014.

J.Urka et al. Censor - a program for identication and elimination of repetitive elements from dna sequences. Computers and Chemistry, 20:119{122, 1996.

Keywords: Scaffolding, Transposable Elements, Repetitive Regions

An -omics search for viruses infecting the phytoparasitic nematode M. incognita

Marc Bailly-Bechet * [1], Carole Belliardo [1], Danchin Etienne [1]

[1] INRA, Université Côte d'Azur, CNRS, Institut Sophia Agrobiotech – Institut national de la recherche agronomique (INRA) – France

Root-knot nematodes (genus Meloidogyne) are phytoparasites and cause important agricultural losses. As for all nematodes, little is known concerning viruses able to infect them: the first nematode viruses were discovered in C. elegans in 2011 (Félix et al, PloS. Biol.), followed by other discoveries among cyst nematodes the same year (Bekal et al. J. Gen. Virol.). In this work, we study in depth the genome and multiple transcriptome datasets of M. incognita, the most devastating representant of the Meloidogyne genus, looking for traces of recent or old viral infections. We first search for sequences of viral origin by homology, with a combination of BLAST and Hidden Markov Models approaches using viral sequences profiles from various publications. Then, to account for sequences of viral origin that may have been dismissed during the genome assembly, we go back to unassembled genomic reads and i) map them to a database of all known viral sequences ii) assemble them before comparing to this database. We finally combine the sequences found by those methods, and filter them when their viral origin is disputable (e.g. ubiquitin sequences). Finding viruses able to infect M. incognita would be both of great agricultural relevance – for the biocontrol perspectives – and of evolutionary relevance, as the study of the parasitism in this genus has shown that key parasitic functions, e.g. entering into the plant root system, are provided by genes acquired by horizontal gene transfer, with only bacteria as donor organisms yet.

Keywords: viruses, nematode, parasite, horizontal gene transfer, homology

*Speaker

Contribution of Oxford NanoPore technology for the de novo genome assembly of non-model species: the case of the Asian ladybird Harmonia axyridis

Jacques Lagnel *† [1], Arnaud Estoup [2], Cécile Donnadieu [3], Céline Lopez-Roques [3], Maxime Manno [3], Anne Loiseau [2], Mathieu Gautier [2]

[1] Unité de recherche Génétique et amélioration des fruits et légumes (GALF) – Institut National de la Recherche Agronomique : UR1052 – France

[2] Centre de Biologie pour la Gestion des Populations (CBGP) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement : UMR55, Centre international d´tudes supérieures en sciences agronomiques : UMR1062, Institut national de la recherche agronomique [Montpellier] : UMR1062, Université de Montpellier : UMR1062, Institut de Recherche pour le Développement : UMR1062, Institut national d'études supérieures agronomiques de Montpellier, Centre international d´tudes supérieures en sciences agronomiques : UMR1062, Centre international d´tudes supérieures en sciences agronomiques : UMR1062, Centre international d´tudes supérieures en sciences agronomiques : UMR1062, Centre international d'études supérieures en sciences agronomiques : UMR1062 – 755 avenue du Campus Agropolis, 34988 Montferrier sur Lez, France

[3] GeT PlaGe, Genotoul – Institut National de la Recherche Agronomique – Campus INRA24 chemin de borde rouge - AuzevilleCS 5262731326 CASTANET-TOLOSAN Cedex, France

The Oxford NanoPore long fragment sequencing echnology (ONT) is based on the sequencing of nucleic acids via the passage of single molecules through protein nanopores and allows sequences of several tens of kilobases to be obtained. It has recently experienced an exponential phase of development and its scope has expanded to many areas of genomics including *de novo*genome assembly.
The ladybug *Harmonia axyridis*from Asia was introduced as a biological control agent but established invasive populations in America, Europe and South Africa.*H. axyridis*has become a model species for the exploration of the genetic determinism of traits involved in the invasive success of some populations. On the other hand, with more than 200 elytra color morphs controlled by a single multi-allelic locus, *H. axyridis*is also an emblematic species for the study of the genetic determinism of color polymorphism. Nevertheless, carrying out the characterization of the genetic architecture of these different traits of interest requires the availability of rich genomic resources and, in particular, good quality genome assemblages (the haploid size of the genome of *H. axyrridis*being estimated at 400 Mb).

Following the relatively disappointing results in terms of assembly quality obtained from conventional Illumina data (sequencing of banks in PE and MP), we generated long sequences from ONT technology (4 flowcells MinIon, 65X, GeT-PlaGe , INRA-Genotoul, France). A *de novo*assembly was realized by combining these long ONT reads with Illumina short reads.

*Speaker

†Corresponding author: jacques.lagnel@inra.fr

Specifically, the assembly pipeline included a correction of the initial reads with the Canu bioinformatics program followed by an assembly with SMARTdenovo. As this assembly had a high level of deletions of homopolymers, a polishing step was carried out with Illumina readings (banks PE and MP). In fine, the contiguity of our assembly is 200 times greater than our initial assembly. Half of the genome obtained is contained in genomic sequences of more than 1.4 Mb with a completeness evaluated by the alignment of conserved orthologous genes (BUSCO) of more than 95.5%. Finally, a contig of 1.3Mb allowed us a fine molecular resolution of a region of the 200 kb genome characterized by a large number of SNPs signifying important molecular variations specific to the different color morphs of *H. axyridis.*

The use of ONT for the *de novo*assembly of our non-model species has been proved an excellent value for money technology.

Keywords: de novo genome assembly, Oxford NanoPore, long fragment sequencing, bioinformatics, insects

Coevolution analysis finds amino-acids involved in immune evasion in Hepatitis B Virus.

Elin Teppa * [1], Francesca Nadalin [1], Alessandra Carbone† [1,2]

[1] Sorbonne Université, CNRS, Laboratory of Computational and Quantitative Biology (LCQB) - UMR 7238 (LCQB) – Centre National de la Recherche Scientifique : UMR7238 – Campus Jussieu Bât. C - 4ème étage 4, place Jussieu 75005 Paris France, France

[2] Institut Universitaire de France (IUF) – Ministère de l'Enseignement Supérieur et de la Recherche Scientifique – France

Introduction

HBV is one of the smallest enveloped DNA viruses and the prototype member of the family Hepadnaviridae. The small HBV genome contains four overlapping open-reading frames (ORF) that encode seven proteins, the viral polymerase (Pol), three surface proteins, two core proteins and the X protein. The gene of the surface proteins consists of a single ORF divided into three in-frame coding regions or domains called preS1, preS2 and S. The large protein (L) comprises the PreS1, preS2 and S domains; the middle surface protein M contains the Pres2 and S domain and the small surface protein S comprises only the small domain. All three HBV surface proteins are integral membrane proteins. They form the main antigen recognized by the immune system, responsible for the attachment of the virus to the hepatocytes and the epitope binding the neutralizing antibodies. Particularly, the S domain contains the major B cell epitope, known as the 'a' determinant. Mutations in and around the 'a' determinant may result in (1) escape of vaccine induced immunity, (2) escaping anti-HBV immunoglobulin therapy and (3) cause diagnostic problems due to false negative results in serological tests (Lazarevic, 2014).
The S ORF is completely overlapped with the polymerase ORF. The polymerase protein comprises four domains named Terminal Protein, Spacer, Reverse Transcriptase (RT) and RNaseH. The RT domain represents the target of the antiviral agents belonging to nucleotide/nucleoside analogues. The currently available drugs for chronic hepatitis B treatment are two immunomodulators and five antiviral agents: lamivudine, telbivudine, entecavir, adefovir and tenofovir (Palumbo, 2008). The major limitation of long-term therapy is antiviral resistance.
The main mutation associated with lamivudine resistance are M204I/V that appear to impair replication, the most common compensatory mutation is L180M, that restores the replicative capacity. *In vitro* studies have demonstrated that this mutation alone is insufficient to result in lamivudine resistance but it augments both viral replication and lamivudine resistance in the context of M204I/V (Bartholomeusz & Locarnini, 2006).
Due to the overlapping between S and Pol ORFs, mutations arising in the RT domain cause the appearance of mutations in the surface proteins conferring to the virus the ability to evade the immune system (Croagh, Desmond, & Bell, 2015; Datta, Chatterjee, Veer, & Chakravarty,

[*] Speaker

[†] Corresponding author: alessandra.carbone@lip6.fr

2012).

A coevolutionary analysis is an attractive approach to find out important positions and predict compensatory mutations associated to the primary resistance mutations. Coevolutionary analysis on HBV sequences represents a challenge due to the high conservation level of protein sequences and the high number of sequences available. Until now, there is no available method to compute coevolution in such data set. In one hand, a large panel of methods exists to compute coevolution in a large set of diverse sequences (de Juan, Pazos, & Valencia, 2013). On the other hand, a handful of methods exist to compute co-evolution signals on small sets of sequences such as CAPS (Fares & McNally, 2006) and BIS2 (Champeimont, Laine, Hu, Penin, & Carbone, 2016; Dib & Carbone, 2012; Oteri, Nadalin, Champeimont, & Carbone, 2017). The BIS2 method was specifically designed to identify clusters of coevolving positions in alignments with high conservation levels (such as viral genomes), or with a relatively low number of sequences (less than 50 sequences). It was successfully applied to reconstruct the protein-protein interaction network of the Hepatitis C Virus (Champeimont et al., 2016) and to identify a novel fusion mechanism in HCV (Douam et al., 2018). However, a novel strategy was needed to compute coevolution using BIS2 in all the available HBV protein sequences.

Here, we present a coevolution analysis of the Pol and L proteins of HBV by applying BIS2 iteratively on selected subsets of the set of HBV sequences, where the set is defined for each viral genotype separately. In the RT Pol domain we found high coevolution signals at positions involved in drug resistance. In the surface protein we found high coevolution in 6 out of 7 positions involved in vaccine escape mutations; in 6 out of 15 involved in immune globulin escape; 3 out of 4 known to be connected with Lamivudine resistance and 10 out of 19 reported as "diagnostic escape" mutants.

Understanding the relationship between mutations in Pol and L proteins may provide valuable information to improve diagnostic procedures and to create most efficient therapeutic protocols.

Material and Methods

Data set

Sequences of L and Pol proteins from genotypes A, B, C and D were retrieved from HBVdb (Hayer et al., 2013). We filtered out incomplete sequences (i.e. truncated proteins) and we retained entries when both protein sequences were available in the same genome. We ended up with 972, 1809, 2006 and 955 sequences belonging to genotypes A, B, C and D respectively. The average of identity in the data set is _˜96% for both proteins. A multiple sequence alignment was built for each genotype for L and Pol proteins using clustal omega (Sievers & Higgins, 2018). The resulting 8 alignments were used as input for the coevolution analysis.

Coevolution method

An iterative strategy was used to compute the BIS2 method in a large number of highly conserved sequences. In the first step, the phylogenetic tree T is predicted from the aligned sequences (BioNJ). In the second step, the sequences corresponding to each subtree T' of T are used to perform BIS2 prediction.

In the third step, statistically significant clusters are selected (P-value ≤ 0.005). In this last step, appropriate criteria are applied to remove the redundancy of the coevolution clusters whenever they are drawn from non-disjoint trees, based on the significance and coevolution pattern and on the number of elements in the clusters.

We considered, for further analysis, the 10 coevolving clusters with best p-value for each genotype. The clusters of coevolving residues are sorted in increasing order by p-value and ranked from 1 to 10.

Results and discussion

Top 10 coevolving clusters in the L protein

We found that positions related to immune escape are present in the top 10 coevolving clusters of the L surface protein in the analyzed genotypes. Namely, we identified:
1) position 130 showing a high coevolution signal in genotypes A, B and C. Mutations at position 130 are related to immune globulin and diagnostic escape.
2) 6 out of 7 positions responsible of vaccine escape.
3) 6 out of 15 positions related to immune globulin escape.
4) due to the overlapping between S and P ORFs, there are mutations in the S ORF that emerge in connection with lamivudine resistance. We found high coevolution in 3 out of 4 of that positions.
5) high coevolution signal in 10 of 19 positions related to diagnostic escape.
The complete description of the top 10 coevolving clusters at the L protein mutant positions associated with immuno or diagnostic escape is summarized in additional file Table 1.

Top 10 coevolving clusters in the RT domain of the Pol protein

In the RT Pol domain, the highest signal of coevolution was found at position 204 in all genotypes showing the amino acid variations M204I/V that correspond to the most common drug resistance mutation.
In genotype A residues, L180 and M204 have the highest signal of coevolution, they coevolve with S109 and N248. The coevolution between L180 and M204 was also found more than once among the top 10 clusters of genotypes C and D, and it corresponds to the most common compensatory mutation. This double mutation is known to confer resistance to lamivudine and telbivudine and to reduce the susceptibility to entecavir and adefovir agents.
In genotype C, L180 and M204 appear twice in the top 10 clusters, and they do not coevolve with any other position.
In genotype D, positions 180 and 204 appear in two clusters, respectively comprised of positions 180, 204 and 229, and positions 80, 91, 180, 204, 253, 266 and 315. The variations L229W/V and L80I/V are associated with telbivudine resistance, whereas the remaining positions have not been previously reported as important positions. Also, the mutations L80V/I have been reported as a compensatory mutation of the primary mutation M204V/I that confers lamivudine resistance. It is worth mentioning that the cluster above, containing the variation L80I/V related to lamivudine resistance, also contains seven positions that belong to Spacer domain in Pol (Pol positions: 178, 208, 235, 286, 290, 294 and 301). This result suggests that important variation related to drug resistance could be outside the RT domain.
In genotype B, position 204 coevolves with positions 80 and 830. As mentioned before, covariation of positions 204 and 80 is associated with drug resistance whereas position 830, which belongs to RNaseH Pol domain, has not been previously reported as an important position.
The analysis of the top 10 coevolving clusters at the RT domain mutant positions associated with drug resistance is summarized in additional file Table 2.

Conclusion

Given the amount of data available for HBV, this virus is a good model to study coevolutionary signals due to biological functions in contrast to structural contacts. The high level of conservation of HBV makes the coevolutionary analysis challenging or even impossible for state-of-the-art covariation methods. In this work, we propose an iterative approach for the BIS2 method that is able to predict clusters of coevolving residues at the L and Pol HBV proteins.
Analyzing the L protein we found high coevolution signals at positions that were reported as

responsible of vaccine escape, immune globulin escape and diagnostic escape.

Regarding the Pol protein we found high coevolution at positions responsible of antiviral resistance, including the known compensatory mutations at positions 204 and 180. The coevolving clusters that contain positions related to drug resistance also include other positions that had not been previously reported as important. Further analyses are needed to evaluate the effect of variation on those positions, it is likely that they may play a role in drug resistance or as compensatory mutation to restore viral fitness.

References

Bartholomeusz, A., & Locarnini, S. A. (2006). Antiviral drug resistance: clinical consequences and molecular aspects. Seminars in Liver Disease, 26(2), 162–170.

Champeimont, R., Laine, E., Hu, S.-W., Penin, F., & Carbone, A. (2016). Coevolution analysis of Hepatitis C virus genome to identify the structural and functional dependency network of viral proteins. Scientific Reports, 6, 26401.

Croagh, C. M., Desmond, P. V., & Bell, S. J. (2015). Genotypes and viral variants in chronic hepatitis B: A review of epidemiology and clinical relevance. World Journal of Hepatology, 7(3), 289–303.

Datta, S., Chatterjee, S., Veer, V., & Chakravarty, R. (2012). Molecular biology of the hepatitis B virus for clinicians. Journal of Clinical and Experimental Hepatology, 2(4), 353–365.

de Juan, D., Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. Nature Reviews. Genetics, 14(4), 249–261.

Dib, L., & Carbone, A. (2012). Protein fragments: functional and structural roles of their coevolution networks. PloS One, 7(11), e48124.

Douam, F., Fusil, F., Enguehard, M., Dib, L., Nadalin, F., Schwaller, L., ... Lavillette, D. (2018). A protein coevolution method uncovers critical features of the Hepatitis C Virus fusion mechanism. PLoS Pathogens, 14(3), e1006908.

Fares, M. A., & McNally, D. (2006). CAPS: coevolution analysis using protein sequences. Bioinformatics, 22(22), 2821–2822.

Hayer, J., Jadeau, F., Deléage, G., Kay, A., Zoulim, F., & Combet, C. (2013). HBVdb: a knowledge database for Hepatitis B Virus. Nucleic Acids Research, 41(Database issue), D566–D570.

Lazarevic, I. (2014). Clinical implications of hepatitis B virus mutations: recent advances. World Journal of Gastroenterology: WJG, 20(24), 7653–7664.

Oteri, F., Nadalin, F., Champeimont, R., & Carbone, A. (2017). BIS2Analyzer: a server for coevolution analysis of conserved protein families. Nucleic Acids Research, 45(W1), W307–W314.

Palumbo, E. (2008). New drugs for chronic hepatitis B: a review. American Journal of Therapeutics, 15(2), 167–172.

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. Protein Science: A Publication of the Protein Society, 27(1), 135–145.

Keywords: coevolution, compensatory mutations, HBV, BIS2, drug resistance

Microbiote et santé : impact des nouveaux pipelines bio-informatiques en métagénomique ciblée

Benoit Goutorbe *† [1,2,3], Anne Plauzolles [2], Ghislain Bidaut [1], Philippe Halfon [2,4]

[1] Centre de Recherche en Cancérologie de Marseille (CRCM) – Aix Marseille Université : UM105, Institut Paoli-Calmettes : UMR7258, Institut National de la Santé et de la Recherche Médicale : U1068, Centre National de la Recherche Scientifique : UMR7258 – 27 bd Leï Roure, BP 30005913273 Marseille Cedex 09, France
[2] Laboratoire Eurpopéen – Laboratoire Alphabio – 1 Rue Melchior Guinot 13003 Marseille, France
[3] Institut National des Sciences Appliquées de Lyon (INSA Lyon) – Département Biosciences – 20 Avenue Albert Einstein, 69621 Villeurbanne cedex, France
[4] Hôpital Européen – Fondation Ambroise Paré – 6 rue Désirée Clary, 13003, France

Les liens entre le microbiote et la santé sont aujourd'hui très étudiés, et ce dans le cadre de diverses pathologies comme les maladies auto-immunes (Li et al., 2018) (Yacoub et al., 2018), les maladies neurodégénératives (Parashar et al., 2017) ou encore en cancérologie (Schwabe et al., 2013). Les études cliniques portant sur le microbiote intestinal se multiplient, permettent d'identifier des signatures caractéristiques de ces pathologies mais les résultats divergent souvent selon les études. Depuis l'avènement des nouvelles technologies de séquençage (NGS, Next Generation Sequencing), la **métagénomique ciblée** de l'ARN 16S est une technique très répandue pour l'étude des communautés bactériennes et un grand nombre d'outils bio-informatiques ont été développés pour traiter les données générées par cette technologie (Edgar, 2013) (Allali et al., 2017) (Caporaso et al., 2010).
Objectifs

Au-delà de la complexité de la question biologique, le manque de standardisation du protocole pré-analytique ainsi que le traitement bio-informatique des données introduisent des biais qui limitent notre compréhension de cet écosystème complexe. Aussi, il est important de comparer les différents outils afin de proposer un cadre d'analyse fiable et uniforme.

Méthodes

Nous appuyant sur des **données réelles** issues d'échantillons de selles ainsi que sur des **données simulées**, nous avons cherché à évaluer et comparer différentes pipelines mis à disposition par la communauté scientifique en isolant les trois étapes majeures du traitement des données primaires : le débruitage, l'assignation taxonomique et la normalisation des tables de comptage.

*Speaker
†Corresponding author: b.goutorbe@alphabio.fr

Des innovations algorithmiques ont été apportées récemment remplaçant le concept d'**OTU** (*Operationnal Taxonomic Unit*) (Edgar et al., 2013) par celui d'**ASV** (*Amplicon Sequence Variants*) (Callahan et al., 2016) qui permet de débruiter les données de manière beaucoup plus fine que ce qui était fait précédemment. Cette nouvelle famille d'algorithme a été très rapidement adoptée sans qu'il existe à ce jour **d'étude comparative objective**.

Lors de l'assignation taxonomique, le choix de l'**algorithme** (classificateurs na´ifs bayésiens sur *Qiime, VSearch, Kraken*) comme celui de la **base de données** (*Silva, Greengenes, Refseq Targeted Loci Project*) utilisée comme référence impacte grandement le résultat final. Nous avons pu évaluer la **performance** des solutions les plus utilisées par la communauté.

Le problème de la normalisation des tables de comptage est commun à un grand nombre de domaines de la bio-informatique qui génèrent des **données compositionnelles**, que ce soit dans l'étude du transcriptome par RNA-Seq ou en cytométrie multiparamétrique. Certaines méthodes issues de ces champs d'application sont donc utilisées (Gloor and Reid, 2016) (Weiss et al., 2017) mais plusieurs aspects rendent le problème particulièrement complexe dans le cas présent. Premièrement, les profils microbiotiques présentent une **très forte diversité**, y compris chez les patients sains, et nous n'avons aucun *a priori* sur des espèces présentes en proportion stable pouvant permettre de calibrer la normalisation. De plus, la **démarche exploratoire** des études actuelles requiert de s'intéresser autant aux populations rares qu'aux populations abondantes. Enfin, pour permettre une analyse multivariée valable, nous nous sommes intéressés à la capacité des différentes méthodes à fournir une mesure de distance **invariante d'échelle**, c'est-à-dire qui permettent aux populations de contribuer équitablement à la distance entre échantillons, quelle que soit leur abondance relative.

Références

Li, B., Selmi, C., Tang, R., Gershwin, M.E., and Ma, X. (2018). The microbiome and autoimmunity: a paradigm from the gut-liver axis. Cell. Mol. Immunol.

Yacoub, R., Jacob, A., Wlaschin, J., McGregor, M., Quigg, R.J., Alexander, J.J. (2018). Lupus: The microbiome angle. Immunobiology *223, 460-465.*

Parashar, A., Udayaba, M.(2017). Gut microbiota: Implications in Parkinson's disease. Parkinsonism & Related Disorders, *38, 1-7.*

Schwabe, R.F., and Jobin, C. (2013). The microbiome and cancer. Nat. Rev. Cancer *13*, 800–812.

Edgar, R.C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Methods *10*, 996–998.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods *13*, 581–583.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Wood, D.E., Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology, *15:46*.

Rognes, T., Flour,i T., Nichols, B., Quince, C., Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ .

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Gl´ockner , F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Opens external link in new windowNucl. Acids Res. *590-596.*

T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie1, K. Keller, T. Huber, D. Dalevi, P. Hu and G. L. Andersen (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Applied and environmental microbiology *72.*

Allali, I., Arnold, J.W., Roach, J., Cadenas, M.B., Butz, N., Hassan, H.M., Koci, M., Ballou, A., Mendoza, M., Ali, R., et al. (2017). A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. BMC Microbiol. *17*, 194.

Gloor, G.B., and Reid, G. (2016). Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can. J. Microbiol. *62*, 692–703.
Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E.R., Knight, R. (2017)**.** Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome *7.*

Keywords: Targeted metagenomics, microbiota, 16S metagenomics

GENOTOUL BIOINFO PLATFORM

Claire Hoede * [1], Céline Noirot *

[1], Jérôme Mariette [1], Christine Gaspin [1], Christophe Klopp [1], Marie-Stephane Trotard [1], Didier Laborie [1], Floreal Cabanettes [1]

[1] Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT INRA) – Institut National de la Recherche Agronomique : UR875 – Chemin de Borde Rouge, 31320 Castanet Tolosan, France

GENOTOUL BIOINFO PLATFORM

Floréal Cabanettes, Christine Gaspin, Claire Hoede, Christophe Klopp, Didier Laborie, Jérôme Mariette, Céline Noirot, Marie-Stéphane Trotard

Introduction

The **GenoToul bioinformatics** (http://bioinfo.genotoul.fr/) facility, created in 2000, provides access to high-performance computing resources, data analysis and programming expertise. During the past 10 years, the team (composed of **7 permanent staff members**) has built up a robust expertise in diverse applications of sequence analysis. This knowledge has been used in software and workflow developments as well as in data analysis projects in various thematics: **genome assembling** (short and long reads), **annotation** (coding and non coding), **sRNA-seq**, **RNA-seq**, **methyl-seq**, **variant analyses**, **metagenomics** (metabarcoding and whole genome) and, more recently, **data integration**.
GenoToul Bioinfo is an INRA strategic platform and is part of both french bioinformatics infrastructure (IFB " Institut Français de Bioinformatique ") and the Genotoul facility (www.genotoul.fr) established in 2000. The platform is located in Toulouse, in the Occitanie region and is hosted by the applied mathematics and informatics laboratory in Toulouse, MIAT.

Main activities :

To provide scaled equipment to biologist and bioinformaticians
Computational resources include a cluster of more than **3000 cores and a storage capacity of 2.5 Po**. Secure access (ssh) is provided to users (_˜1200 users at the end of 2017). Equipments are hosted at the INRA datacenter located in Auzeville, south of Toulouse, providing a high security level, such as electric power, temperature, fire protection, intrusion controls. Redundant equipments and daily back-ups also insure data protection and security. All these

*Speaker

equipments are open to the life science community requiring bioinformatics resources.

To set up an environment with software and databanks

More than **230 databanks** are regularly updated, **900 bioinformatic software** were installed by the end of 2017. Furthermore, the platform hosts a Galaxy instance deployed and maintained by the Sigenae team (http://www.sigenae.org/).
To train biologists (and bioinformaticians)
With **Sigenae**, **NED** (GenPhySE), **SaAB** (MIAT) and **TWB**, the GenoToul bioinfo platform contributes to train biologists and bioinformaticians. The training catalog and registration forms are available on the website. In 2017, the platform organized 20 training days.

To support biologist teams for data analysis

Team members provide supports to biologists to analyse their data, mainly HTS (high throughput sequencing technologies) sequences analysis. In 2017, the platform was involved in 26 projects and has co-signed 17 publications with biologists.

To develop Software for the community

The team develops novel data analysis methods and tools. The most recent tools include an interactive dot plot viewer adapted to large genomes (http://dgenies.toulouse.inra.fr/) and methods to integrate heterogeneous data. The team also sets up or maintains data analysis workflows. In 2017, three articles were published in collaboration with the Sigenae and SaAB teams.

GenoToul Bioinfo platform BioInfo GenoToul has a quality management system covering all of its activities and is certified since 2010 (NFX50-900 and ISO 9001).

Keywords: genotoul, cluster, software, databank, training, data analysis, software development

Transcriptomics analysis using Long Read sequencing

Rachel Legendre *† [1,2], Juliana Pipoli Da Fonseca [2], Thomas Cokelaer‡ [1,2]

[1] Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 26-28 Rue du Docteur Roux, 75015 Paris, France
[2] Biomics – Institut Pasteur de Paris – 28 rue du Docteur Roux 75015 Paris, France

Identification of complete transcriptome remains a big challenge. Indeed, despite well-established techniques based on short-read technologies, precise identification of isoforms or full-length transcripts remains a difficult task. Thanks to long-read technologies it is now possible to sequence complete isoforms. For instance, the Sequel System of PacBio (PacBio Biosciences) provides the so-called IsoSeq method, which allows users to generate full-length cDNA of very high quality; that way, the entire transcriptome can be confidently characterized.
In this work, we present preliminary results regarding an IsoSeq experiment, which aimed at improving current annotations of 3 different yeast strains. In this experiment, PacBio sequencing was performed using the SMARTer PCR cDNA Synthesis Kit. SIRV (Spike-In RNA Variant Control Mixes) were also injected in all samples in order to control variability and efficiency of library sequencing. Finally, an Oxford Nanopore Technologies (ONT) on the same strain was performed to allow a PacBio/ONT comparison.
First, we show that we can identify the injected SPIKES in the consensus reads (CCS hereafter). Second, we analyzed the raw reads with the SMRTlink software in order to obtain high quality (HQ) isoforms. The HQ isoforms allow us to identify known genes with precise 5' and 3' boundaries. Moreover, with CCS reads, we show that the IsoSeq method is reproducible and allow us to detect the full set of transcripts. Finally, we compare the transcripts detected with the Nanopore and PacBio technologies: we show that the overlap between the two methods is high (99%).

Keywords: PacBio, Isoseq, transcriptomics, Long Reads, Nanopore

*Speaker

†Corresponding author: rachel.legendre@pasteur.fr

‡Corresponding author: thomas.cokelaer@pasteur.fr

The Migale bioinformatics platform

Valentin Loux *† [1], Sam Ah Lone [1], Maxime Branger‡ [1,2], Hélène Chiapello [1], David Christiany [1,3], Sandra Derozier [1], Olivier Inizan [1], Valentin Marcon [1], Véronique Martin [1], Cédric Midoux [1,4], Eric Montaudon [1], Thi-Phuong-Lien Nguyen [1,3], Olivier Rué [1], Sophie Schbath [1], Valérie Vidal [1]

[1] MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France (MaiAGE) – Institut national de la recherche agronomique (INRA) : UR1404 – Batiment 233, centre de Vilvert, 78350 Jouy en Josas CEDEX, France
[2] Infectiologie Santé Publique (ISP-311) – Institut National de la Recherche Agronomique : UMR1282, Université de Tours – 37380 Nouzilly, France
[3] Institut de Biosciences et Biotechnologies de Grenoble (BIG) – Université de Grenoble-Alpes, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble, Centre National de la Recherche Scientifique : FR3425 – 17, rue des Martyrs 38054 Grenoble cedex 9, France
[4] Hydrosystems and Bioprocesses Research Unit, Irstea, France (HBAN) – Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture - IRSTEA – 1 rue Pierre-Gilles de Gennes 92761 Antony Cedex, France

The Migale bioinformatics platform is a team of the INRA's MaIAGE research unit (Applied Mathematics and Computer Science, from Genome to the Environment).It has been existing since 2003 and is intended to provide services to the life sciences community.
The Migale platform offers four types of services:

- an open infrastructure dedicated to life sciences data processing,
- dissemination of expertise in bioinformatics,
- design and development of bioinformatics applications,
- data analysis.

Migale is part of the French Institute of Bioinformatics (IFB) and France Génomique projects.

The poster will illustrate the platform's missions and offered services with examples chosen from recent achievements: training cycle "bioinformatics through practice", development of a database of microbial phenotypes, metagenomic data analysis service and projects.
http://migale.jouy.inra.fr

*Speaker
†Corresponding author: valentin.loux@inra.fr
‡Corresponding author: maxime.branger@inra.fr

Keywords: bionformatics platform, metagenomic analysis, software development

BIPAA, Bioinformatics Platform for the Agroecosystems Arthropods.

Stéphanie Robin *† [1,2], Anthony Bretaudeau [1,2], Fabrice Legeai [1,2]

[1] Institut de Génétique, Environnement et Protection des Plantes (IGEPP) – Institut National de la Recherche Agronomique : UMR1349, Universite de Rennes 1 : UMR1349, Agrocampus Ouest : UMR1349 – Domaine de la Motte au Vicomte BP 3532735653 Le Rheu, France

[2] Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) – Universite de Rennes 1, Institut National des Sciences Appliquées - Rennes, Université de Bretagne Sud, École normale supérieure - Rennes, Institut National de Recherche en Informatique et en Automatique, CentraleSupélec, Centre National de la Recherche Scientifique : UMR6074, IMT Atlantique Bretagne-Pays de la Loire – Avenue du général LeclercCampus de Beaulieu 35042 RENNES CEDEX, France

BIPAA (BioInformatics Platform for the Agroecosystems Arthopods) (https://bipaa.genouest.org) is a bioinformatics platform from the French National Institute for Agricultural Research (INRA). It is located in Rennes (France) and is integrated in the GenOuest infrastructure (https://www.genouest.org). It is dedicated to assist genomics and post-genomics programs developed on insects associated to agroecosystems. More than six hundred users are currently listed on BIPAA.
#Data analysis. BIPAA is supporting a network of scientists from various french labs for analyzing their genomics data. Depending on the needs, it implies the personal guidance for developping scripts, running complex workflows on a computing cluster or on Galaxy servers. We collaborate with many biology labs for computing and analyzing heterogeneous data covering many bioinformatics topics such as genome assembly and annotation, expression analysis, non-coding RNA characterization, genomes comparison, variant identification, or genomics and epigenomics data integration.

#Databases. BIPAA is the home of several public reference databases hosting multiple insect genomes: AphidBase (for aphids), LepidoDB (for lepidopterans) and ParWaspDB (for parasito'id wasps). 21 genomes are currently available on line. For each genome, a collection of web applications allow to explore reference genome or transcriptome assemblies and annotations (e.g. genome browser, gene reports), to analyze this data (e.g. dedicated Galaxy server, specific web applications), and to collect new scientific knowledge (e.g. manual curation of annotations using Apollo).

#Collaborations. Often in collaboration with the GenScale and Dyliss teams in INRIA/Irisa in Rennes (France), BIPAA is engaged in various research programs involving bioinformatics skills. For example, an user-friendly web application for integrating and querying heterogeneous data (AskOmics) or a tool for long non-coding RNA annotation (FEELnc) were developped in this context. BIPAA is also associated with 2 national networks of INRA : BAPOA (Biologie Adaptative des Pucerons et Organismes Associés) and ADALEP (Adaptation à l'environnement

*Speaker

†Corresponding author: stephanie.robin@inra.fr

biotique chez les lépidoptères) networks. It is also involved in international consortia or various insect genome sequencing projects like i5k, an initiative to sequence the genomes of 5000 arthropods.

#Trainings. Moreover, frequent training sessions in partnership are organized with the GenOuest bioinformatics platform or the BBRIC (Bioinformatique, Biodiversité, Représentation et Intégration des Connaissances) community of INRA.

Keywords: Platform, genomics, insects, agroecosystems

Sequanix: a dynamic graphical interface for Snakemake workflows

Thomas Cokelaer *† [1], Dimitri Desvillechabrol [2]

[1] Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 26-28 Rue du Docteur Roux, 75015 Paris, France
[2] Institut Curie, PSL Research University, F-75005 Paris, France – Institut Curie, PSL Research University – France

We designed a PyQt graphical user interface - Sequanix - aimed at democratizing the use of Snakemake pipelines in the NGS space and beyond. By default, Sequanix includes Sequana NGS pipelines (Snakemake format) (http://sequana.readthedocs.io), and is also capable of loading any external Snakemake pipeline. New users can easily, visually, edit configuration files of expert-validated pipelines and can interactively execute these production-ready workflows. Sequanix will be useful to both Snakemake developers in exposing their pipelines and to a wide audience of users.

Availability and implementation: the source code is available on http://github.com/sequana/sequana, bio-containers on http://bioconda.github.io and Singularity hub (http://singularity-hub.org).

Keywords: snakemake, GUI, NGS

*Speaker

†Corresponding author: thomas.cokelaer@pasteur.fr

Bioconvert: a collaborative bioinformatics format converter library.

Anne Biton [1], Bryan Brancotte [1], Thomas Cokelaer *† [1], Yoann Dufresne [1], Kenzo-Hugo Hillion [1], Etienne Kornobis [1], Pierre Lechat [1], Rachel Legendre [1], Frédéric Lemoine [1,2], Blaise Li [1], Nicolas Maillet [1], Bertrand Néron [1], Amandine Perrin [1], Rachel Torchet [1], Nicolas Traut [3], Anna Zhukova [1,2]

[1] Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 26-28 Rue du Docteur Roux, 75015 Paris, France

[2] Unité Bioinformatique Evolutive, C3BI USR 3756, Institut Pasteur CNRS – Institut Pasteur de Paris – Paris, France

[3] Unité de Génétique Humaine et Fonctions Cognitives, Département de Neuroscience, Institut Pasteur – Institut Pasteur de Paris – Paris, France

Life sciences involve the knowledge and use of many different data formats. Their diversity, complexities and the lack of appropriate tools may lead to cumbersome and sometimes challenging conversions between these formats. Many conversion tools already exist but exhibit drawbacks such as being dispersed, focused on a few specific formats, difficult to install, not optimised or obsolete. With Bioconvert, we plan to cover a wide spectrum of format conversions in a single entry point. To do so, we design a simple framework that allows to use existing tools when available and to implement new conversions when they do not exist. We also implemented a benchmark framework to ease the comparison between different conversion tools.
Bioconvert project has only recently started (at the end of 2017), nevertheless, thanks to a collaborative approach, there are already about 80 conversions available, including 40 different formats. Each conversion may have different implementations leading to about 120 unique methods.
Bioconvert is available on github at https://github.com/biokit/bioconvert and from pypi website. The project follows modern software development and good practices including openness, testing, continuous integration and automatic online documentation. Documentation is available at: http://bioconvert.readthedocs.io . In addition to a standalone version available from source or via existing bio-containers (on bioconda or as Singularity image), we also plan to provide an online version where users can easily convert their data without having to install the software locally.

Keywords: bioconvert, NGS formats, singularity, web application, conversion, python

*Speaker

†Corresponding author: thomas.cokelaer@pasteur.fr

Novel approaches for phylogenetic analysis of NGS data

Yannick Antoine * [1,2], Benjamin Linard† [2], Krister Swenson [2], Fabio Pardi‡ [2]

[1] Master BCD - Université de Montpellier – Université Montpellier II - Sciences et Techniques du Languedoc – 163 rue Auguste Broussonnet 34090 Montpellier, France
[2] Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier : UMR5506, Centre National de la Recherche Scientifique : UMR5506 – 161 rue Ada - 34095 Montpellier, France

Species identification remains a complex problem when it comes to complex samples such as metagenomes as they generally hold unknown clades. Likelihood-based phylogenetic inference is generally considered as the most accurate approach. Its application on millions of query fragments is technically impossible with current implementations. Phylogenetic placement PP algorithms were developed in the context of prokaryote metagenomics as an alternative to phylogenetic inference. In particular, phylogenetic placement avoids inter query Likelihood computations and focuses only the assignation of query sequences to a particular branch of a fixed reference phylogeny. In practice, the user provides a reference alignment and a phylogeny (built from this alignment). Then, query sequences are aligned to the reference alignment, a step done generally through HMM-based alignments. The resulting matrix is analyzed by the phylogenetic placement algorithm to independently place each query sequence on other tree branches maximizing the Likelihood of the corresponding insertion. While robust in determining the taxonomic composition of metagenomes (Matsen et al., 2011), previous phylogenetic placement algorithms struggle to follow the increasing throughput of sequencing technologies.

To improve the scalability of phylogenetic placement, RAPPAS (Rapid Alignment-free Phylogenetic Placement via Ancestral Sequences) was developed, a new algorithm of phylogenetic placement designed to reduce the computational cost of phylogenetic placement. The process of RAPPAS is split in two main phases: database construction and query placement. The first phase is based on standard ancestral sequence reconstruction (ASR) techniques which are applied on the reference tree and alignment to build ancestral k-mers associated to each branch of the reference tree, i.e. a collection of short words corresponding to divergence scenarios which may have occurred on each of these branches. Our expectation is that future query sequence may match the ancestral k-mers generated for a specific branch and validate this scenario. The most relevant fraction of this search space is then stored in an ancestral k-mers database (noted akmDB), in which ancestral k-mers are associated to the branch from which they were generated and the associated posterior probabilities. Later, the user will exploit the akmDB to produce the phylogenetic placement which are linear to the query length (no alignments are involved,

*Speaker

†Corresponding author: benjamin.linard@lirmm.fr

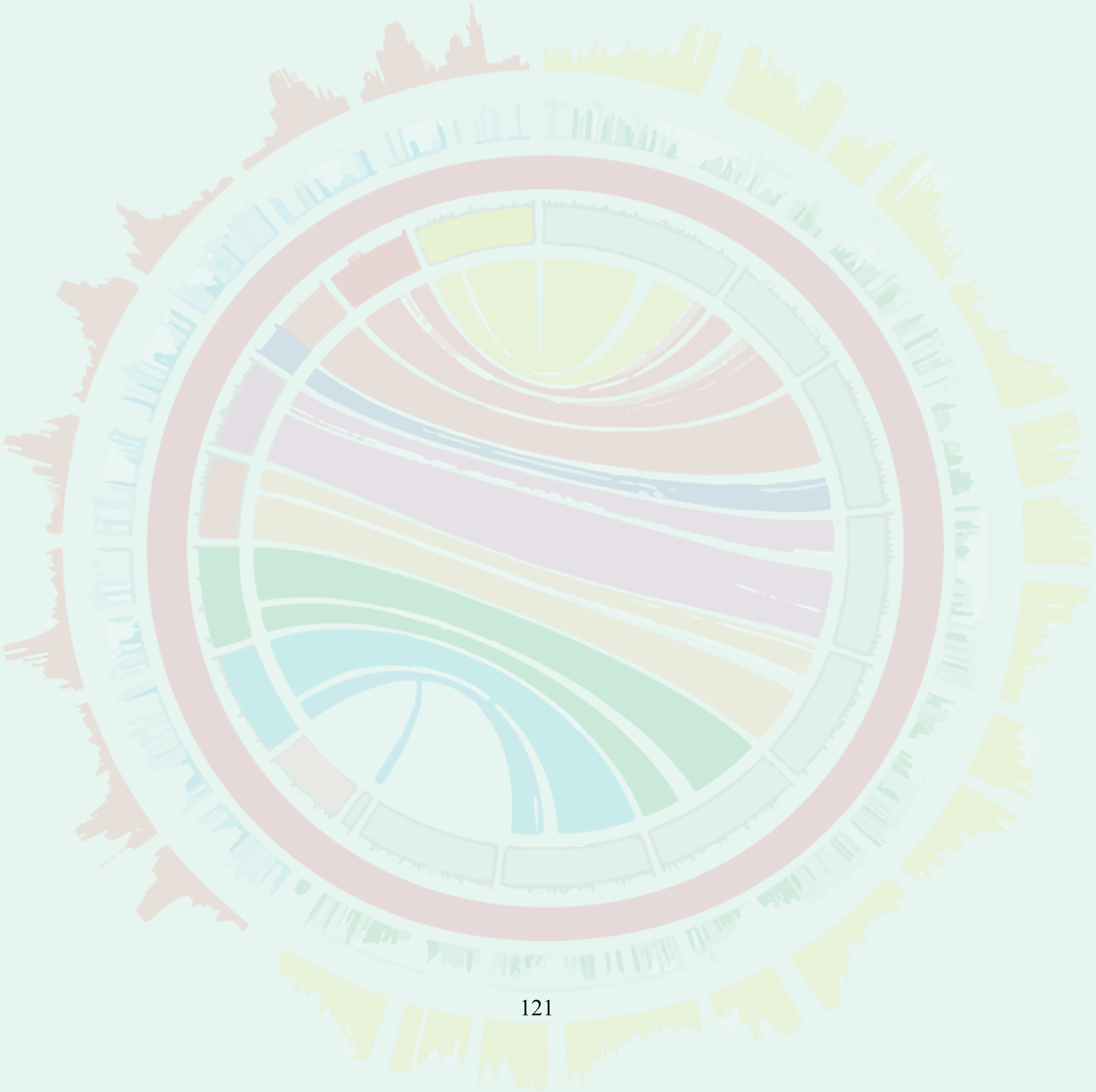
‡Corresponding author: fabio.pardi@lirmm.fr

only k-mer matches). RAPPAS is available at https://gite.lirmm.fr/linard/RAPPAS and is today the fastest algorithm of phylogenetic placement, while simultaneously keeping a placement accuracy which is similar to previous algorithms.

Currently, this alignment-free approach is fast but do not offer a mean to select which metagenomic reads should be indeed placed on the reference tree. Indeed, if a metagenomic read is not linked to the reference, it will still be placed despite a very low score. Previous methods were in fact based on the preliminary step of alignment to select which read should be aligned (using alignment score thresholds), a common approach being to use HMM profile alignments to make a selection of reads to place. Therefore, our objective is to introduce in RAPPAS a measure estimating the degree of homology linking query to the k-mer database, to provide phylogenetic placement results only for the metagenomic reads which are homolog to the sequences of the reference tree into which they are placed. These developments involve the development of new scores based on the collinearity of the k-mers matches between query and reference alignment, as well as the associated probabilites. I will will introduce the algorithm of RAPPAS as well as this new extensions.

Keywords: Metagenomic, Phylogenetic placement, species identification, ancestral sequence reconstruction

New challenges for bioinformatics in the personalized medicine era



Identification de marques épigénétiques marqueurs de prédisposition aux maladies métaboliques

Jérémy Tournayre *† [1], Laurent Parry [1], Anne-Catherine Maurin [1], Julien Averous [1], Alain Bruhat [1], Valerie Carraro [1], Aurore Douge [1], Cyrielle Vituret [1], Celine Jousse‡ [1], Pierre Fafournoux§ [1]

[1] Unité de Nutrition Humaine - Clermont Auvergne (UNH) – Université Clermont Auvergne : UMR1019, Institut national de la recherche agronomique [Auvergne/Rhône-Alpes] : UMR1019 – Centre de Recherches Inra de Clermont-Fd/Theix / 63122 St Genès Champanelle, France

Contexte

Les causes du développement de nombreuses maladies métaboliques telles que le diabète et l'obésité sont multifactorielles et peuvent inclure une prédisposition génétique, mais aussi l'influence de facteurs environnementaux. Parmi ces facteurs environnementaux, l'alimentation et en particulier l'alimentation pendant les phases précoces de la vie joue un rôle particulièrement important. En effet, de nombreuses études épidémiologiques montrent qu'une mauvaise nutrition chez la femme enceinte a des répercussions importantes sur la santé de l'enfant tout au long de sa vie. Ce stress nutritionnel maternel affecte un grand nombre de paramètres métaboliques liés au Syndrome Métabolique chez les descendants adultes. Ces travaux ont fait émerger le concept de programmation nutritionnelle fœtale se traduisant par la mise en place d'empreintes laissées par l'environnement maternel initial et qui persistent tout au long de la vie. Les modifications épigénétiques jouent un rôle important dans ce processus. Le terme épigénétique décrit des altérations stables de l'expression des gènes qui n'impliquent aucun changement de la séquence nucléotidique de l'ADN. Ce type de régulation survient au cours du développement et/ou en fonction de l'environnement et se maintient au travers de la mitose. Parmi les différentes marques épigénétiques existantes, nous avons choisi de nous intéresser à la méthylation de l'ADN.
Le modèle murin d'imprégnation nutritionnelle fœtale utilisé au laboratoire consiste à donner un régime Low Protein Diet (LPD) (contenant 10% de protéine) à des femelles Balb/c pendant les périodes de Gestation et/ou de Lactation. Les descendants mâles (F1) obtenus sont tous nourris avec un régime contrôle à partir du sevrage à 1 mois. A l'âge adulte, les descendants issus des mères ayant reçu un régime LPD pendant la gestation et la lactation (F1-LPDGest+Lact) sont résistants à la prise de poids induite par la consommation d'un régime obésogène, sont hypermétaboliques et présentent les caractéristiques d'individus résistants au Syndrome Métabolique (Jousse et al., 2014). A l'inverse, les descendants issus des mères ayant reçu un régime LPD pendant la gestation seule (F1-LPDGest) présentent un phénotype inverse et sont sensibles à la prise de poids induite par la consommation d'un régime obésogène. Ils présentent donc les caractéristiques d'individus sensibles au Syndrome Métabolique.

*Speaker

†Corresponding author: jeremy.tournayre@inra.fr

‡Corresponding author: celine.jousse@inra.fr

§Corresponding author: pierre.fafournoux@inra.fr

Notre objectif est d'identifier, chez les descendants dont les mères ont été soumises à divers stress nutritionnels, les régions différentiellement méthylées (Differentially Methylated Region = DMR) qui corrèlent de façon spécifique avec le phénotype résistant ou sensible au Syndrome Métabolique. Ces DMR pourraient être considérés comme marqueurs prédictifs d'une sensibilité ou résistance au Syndrome Métabolique et utilisables comme outil diagnostique ou comme cible thérapeutique.

Nous présentons ici la stratégie bio-informatique développée pour l'analyse des données obtenues.

Whole Genome Bisulfite Sequencing (WGBS)

A partir du modèle murin décrit ci-dessus, le tissu adipeux blanc périgonadal des souris F1 adultes de chacun de nos groupes expérimentaux a été prélevé afin d'en extraire l'ADN génomique. La première technique employée pour l'identification des régions différentiellement méthylées (DMR) est le Whole Genome Bisulfite Sequencing qui permet d'identifier des DMR sans a priori à partir d'ADN génomique traité au bisulfite. Le traitement bisulfite induit une modification chimique qui transforme en Thymine (T) toutes les Cytosines (C) du génome à l'exception des C méthylées. Ainsi, ce traitement permet de transformer l'information C méthylée / C non méthylée en polymorphisme C / T qui peut être mesuré de façon quantitative par séquençage. Dans un premier temps un mélange d'échantillons extraits du tissu adipeux blanc périgonadal pour chaque phénotype (contrôle, résistant et sensible au Syndrome Métabolique) a été utilisé. La détection de DMR a nécessité la construction d'une méthode informatique basée sur le principe des fenêtres coulissantes. 5000 DMR issus de ces données ont été détectées par cet outil.

Targeted bisulfite sequencing : épicapture (seq-Cap Epi Enrichment / Roche®)

Les DMR identifiés par la technique décrite au paragraphe précédent ont par la suite été analysés par une autre technique : l'épicapture. Brièvement, à partir d'ADNg traité au bisulfite, l'épicapture permet la capture puis le séquençage de régions d'intérêt grâce à des sondes spécifiques.
Cette technique, qui a un coût moins élevé que le WGBS, nous a permis de cibler des régions d'intérêt sur un nombre d'échantillons suffisant dans tous nos groupes expérimentaux.

Nous avons mis en place le traitement des données issues de l'épicapture. Ce traitement consiste en plusieurs étapes 1) traitement qualité des données avec Trimmomatic (Bolger et al., 2014), 2) alignement sur le génome de la souris (version mm9) avec Bismark (Krueger and Andrews, 2011), 3) filtrage sur la couverture, 4) identification des DMR avec l'outil Metilene (J´uhling et al., 2015).

Identification des Differentially Methylated Region

On a défini deux listes de DMR :
- une liste de DMR avec des paramètres de stringence faible afin d'avoir une liste la plus large possible qui permettra de réaliser des épicaptures dans le même modèle expérimental mais sur d'autres tissus (Foie, Muscle, PBMC...) ;
- une liste de DMR robustes avec des paramètres de stringence classique (q-value < =5%) afin de réaliser une analyse fonctionnelle.

On a caractérisé les DMR à partir de la liste robuste. Chaque DMR peut être relié au phénotype résistant au Syndrome Métabolique et/ou au phénotype sensible au Syndrome Métabolique. Afin d'identifier si un DMR est spécifique d'un phénotype, nous avons mis en place une méthode informatique qui prend en compte le fait que les DMR identifiés puissent ne pas être à des positions parfaitement identiques entre les deux phénotypes.

Annotations des DMR

Nous choisissons d'attribuer un gène à un DMR si celui-ci chevauche/est compris entre -2 kb du site d'initiation de la transcription d'un gène et +2 kb du site de terminaison de la transcription d'un gène. Ces DMR qui sont liés à des gènes proches sont appelés des DMG (Differentially Methylated Genes). Pour cela, trois bases de données d'annotation génomique sont choisies : RefSeq (Pruitt et al., 2007), Ensembl (Zerbino et al., 2018) et UCSC (Casper et al., 2018). De la même manière que pour les DMR, les DMG sont classés selon leur spécificité vis-à-vis du phénotype : sensibilité et/ou résistance au Syndrome Métabolique.

Afin de savoir si nos DMR peuvent être dans des régions accessibles de la chromatine et s'ils possèdent des sites de fixation de facteurs de transcription nous avons créé un outil qui à partir d'une région chromosomique récupère les données de ChipSeq, d'ATAC-seq et de Dnase-seq identifiées dans d'autres expériences à l'aide des bases de données CistromeDB (Mei et al., 2017), GTRD (Yevshin et al., 2017) et de la base de données de motifs de facteurs de transcription Jaspar (Mathelier et al., 2014).

Identification des Differentially Expressed Genes (DEG)

Pour corréler nos données de méthylation à des valeurs d'expression génique, une analyse RNA-seq a été réalisée à partir d'ARN extrait des mêmes échantillons de tissus adipeux blanc périgonadal. Le traitement informatique de ces données est réalisé à partir des outils disponibles sur " usegalaxy.org " (Afgan et al., 2016) : filtration des données selon leur qualité, alignement sur la version mm9 du génome de la souris, filtration sur la couverture, assemblage et comparaison de chaque groupe expérimental par rapport au groupe contrôle avec l'outil CuffDiff (Trapnell et al., 2013) sur la version mm9 du transcriptome de référence de la souris de RefSeq. De même que pour les DMR et les DMG, les DEG spécifiques des phénotypes résistant et/ou sensible au Syndrome Métabolique ont été identifiés.

L'analyse en Gene Ontology (GO) de la liste des DEG liés au phénotype résistant au Syndrome Métabolique avec l'outil ProteINSIDE (Kaspric et al., 2015) donne des GO attendus de même que l'analyse de la liste des DEG liés au phénotype sensible au Syndrome Métabolique.

Conclusion

Nous avons obtenu une liste de régions différentiellement méthylées spécifiques de chacun de nos phénotypes (résistant ou sensible au Syndrome Métabolique). Notre objectif étant d'identifier des marqueurs prédictifs d'une sensibilité ou résistance au Syndrome Métabolique utilisables en médecine, les DMR identifiés dans notre modèle animal devront être testés chez l'Homme. Nous commencerons par identifier les régions orthologues avec l'outil LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Les DMR ainsi transposés à l'Homme pourront être testés par épicapture sur des échantillons humains issus de cohortes caractérisés pour leurs paramètres métaboliques. Le pipeline bio-informatique décrit précédemment est prêt à être utilisé pour l'analyse des résultats issus de ces prochaines épicaptures chez l'Homme.

Bibliographie

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D., et al. (2018). The UCSC Genome Browser database: 2018 update. Nucleic Acids Res. 46, D762–D769.

Jousse, C., Muranishi, Y., Parry, L., Montaurier, C., Even, P., Launay, J.-M., Carraro, V., Maurin, A.-C., Averous, J., Chaveroux, C., et al. (2014). Perinatal Protein Malnutrition Affects Mitochondrial Function in Adult and Results in a Resistance to High Fat Diet-Induced Obesity. PLOS ONE 9, e104896.

J́uhling, F., Kretzmer, H., Bernhart, S.H., Otto, C., Stadler, P.F., and Hoffmann, S. (2015). metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Res.

Kaspric, N., Picard, B., Reichstadt, M., Tournayre, J., and Bonnet, M. (2015). ProteINSIDE to Easily Investigate Proteomics Data from Ruminants: Application to Mine Proteome of Adipose and Muscle Tissues in Bovine Foetuses. PloS One 10, e0128086.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27, 1571–1572.

Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C., Chou, A., Ienasescu, H., et al. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 42, D142–D147.

Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., Zhu, M., Wu, J., Shi, X., Taing, L., et al. (2017). Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res. 45, D658–D662.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61–D65.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. 31, 46–53.

Yevshin, I., Sharipov, R., Valeev, T., Kel, A., and Kolpakov, F. (2017). GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Res. 45, D61–D67.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res. 46, D754–D761.

Keywords: nutrition, syndrome métabolique, épigénétique, méthylation, transcriptomique, RNAseq

I-LowVarFreq : improving low-frequency variant detection using a new UMI-based variant calling approach for paired-end sequencing NGS libraries.

Pierre-Julien Viailly *† [2,1], Elodie Bohers [2,1], Mathieu Viennot [2,1], Ahmad Abdel Sater [1,2], Hélène Dauchel [3], Thierry Lecroq [3], Pierre Vera [2,3], Fabrice Jardin‡ [1,2]

[2] Centre Henri Becquerel, 76000 Rouen, France – Centre de Lutte Contre le Cancer Henri Becquerel Normandie Rouen – France
[1] Normandie Univ, UNIROUEN, INSERM U1245, Team "Genomics and Biomarkers of Lymphoma and Solid Tumors", 76000 Rouen – INSERM U1245 – France
[3] Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France – UNIHAVRE – France

The study of low frequency variants in a range of biological application has been transformed by recent advances in sequencing throughput. For example, within the field of oncology, liquid biopsy can potentially be used to monitor tumor burden in the blood. Studying tumor heterogeneity or biopsies poor in tumoral cells imply also to detect variants at low frequency level. Low-frequency variants are often confounded by sequencing errors and DNA polymerase errors during PCR enrichment. Unique molecular identifiers (UMI) have been used in recent targeted sequencing protocols to address this issue but the bioinformatics analysis of such data still remain challenging. The UMI are short random molecular tags added to each targeted DNA fragment before library amplification in NGS protocols. They allow to identify duplicated reads after sequencing and also to detect sequencing and PCR errors.

Here, we present I-LowVarFreq, a new UMI variant calling strategy with remarkably higher specificity compared to raw-reads-based variant calling.

UMIs are first extracted from BAM files using UMI-tools.

Then, I-LowVarFreq generates an initial list of candidate variants using SAMTools pileup and Poisson modeling. Aligned reads are reconstructed using CIGAR and raw read sequences. For each variant position, UMI tags were extracted from overlapping reads.

Finally, I-LowVarFreq predicts sequencing artifacts from candidates using UMI counts. UMI counts are divided into 7 classes :

*Speaker

†Corresponding author: pierre-julien.viailly@chb.unicancer.fr

‡Corresponding author: fabrice.jardin@chb.unicancer.fr

UMIwt/mt : number of unique wild-type (1) or mutated (2) UMIs in both reads of read pair

strong UMIwt/mt: number of non-unique wild-type (3) or mutated (4) UMIs in both reads of read pair

singleton UMIwt/mt : number of unique wild-type (5) or mutated (6) UMIs without read pair

discordant UMIs : number of discordant UMIs (7), with dissonant allele call for the same read pair

We illustrate the results obtained using I-LowVarFreq through the sequencing results of several types of lymphoma. Example of matched biopsy and cell-free DNA sequencing results will be discussed and I-LowVarFreq sensitivity/specificity will be compared to other variant calling approaches.

Keywords: UMI, variant calling, low variation frequency, cancer, sequencing artifacts

Perturbed human sub-networks by Fusobacterium nucleatum candidate virulence proteins

Andreas Zanzoni *† [1], Lionel Spinelli [1], Shérazade Braham [1], Christine Brun [1]

[1] Theories and Approaches of Genomic Complexity (TAGC) – Aix Marseille Université, Institut National de la Santé et de la Recherche Médicale : U1090, Centre National de la Recherche Scientifique – Théories et approches de la complexité génomiqueParc scientifique de Luminy163 avenue de Luminy13288 Marseille cedex 9, France

Fusobacterium nucleatum is a gram-negative anaerobic species residing in the oral cavity and implicated in several inflammatory processes in the human body. Although *F. nucleatum* abundance is increased in inflammatory bowel disease subjects and is prevalent in colorectal cancer patients, the causal role of the bacterium in gastrointestinal disorders and the mechanistic details of host cell functions subversion are not fully understood.
How could *F. nucleatum* hijack human cells? Pathogens employ a variety of molecular strategies to reach an advantageous niche for survival. One of them consists of subverting host protein interaction networks. To achieve this, virulence factors often display structures resembling host components in form and function to interact with host proteins, thus providing a benefit to the pathogen. Such "molecular mimics" (e.g., targeting motifs, enzymatic activities, and protein–protein interaction elements) allow pathogens to enter the host cell and perturb cell pathways.

In order to gain new insights on the molecular cross-talk between *F. nucleatum* and the human host, we devised a computational strategy to identify putative secreted *F. nucleatum* proteins (*Fuso*Secretome) and to infer their interactions with human proteins based on the presence of host molecular mimicry elements. We mapped the FusoSecretome interactors on a binary human interactome and, by using the OCG algorithm, we studied its modular structure of in order to identify host cellular functions that are likely perturbed by *F. nucleatum* candidate virulence proteins.

*Fuso*Secretome proteins share similar features with known bacterial virulence factors thereby highlighting their pathogenic potential. We show that they interact with human proteins that participate in infection- related cellular processes and localize in established cellular districts of the host-pathogen interface. Our network-based analysis identified 31 functional modules in the human interactome preferentially targeted by 138 *Fuso*Secretome proteins, among which we selected 26 as main candidate virulence proteins, representing both putative and known virulence proteins. Finally, six of the preferentially targeted functional modules are implicated in the onset and progression of inflammatory bowel diseases and colorectal cancer. Overall, our

*Speaker

†Corresponding author: andreas.zanzoni@univ-amu.fr

computational analysis identified candidate virulence proteins potentially involved in the *F. nucleatum*-human cross-talk in the context of gastrointestinal diseases.

Over the last years, many microbes have been identified as key players in chronic disease onset and progression. However, untangling these complex microbe–disease associations requires lot effort and time, especially in the case of emerging pathogens that are often difficult to manipulate genetically. In this context, our computational strategy can be helpful in guiding and speeding-up wet lab research in microbe–host interactions.

Keywords: network biology, molecular mimicry, short linear motifs, protein interaction inference, colorectal cancer, inflammatory bowel's disease

Sarek, a portable workflow for WGS analysis of germline and somatic mutations

Maxime Garcia *† [1,2,3,4,5], Juhos Szilveszter [1,2,5], Malin Larsson [4,6,7,8], Teresita Díaz De Ståhl [1,2,5], Johanna Sandgren [1,2,5], Jesper Eisfeldt [5,9,10], Sebastian Dilorenzo [4,7,11,12], Marcel Martin [4,7,13,14], Pall Olason [4,7,12,15], Phil Ewels [3,4,13,14], Björn Nystedt [4,7,12,15], Monica Nister [1,2,5], Max Käller [3,4,16,17]

[1] The Swedish Childhood Tumor Biobank (Barntumörbanken) – Sweden
[2] Dept. of Oncology Pathology – Sweden
[3] National Genomics Infrastructure (NGI) – Sweden
[4] Science for Life Laboratory (SciLifeLab) – Sweden
[5] Karolinska Institutet – Sweden
[6] Dept. of Physics, Chemistry and Biology – Sweden
[7] National Bioinformatics Infrastructure Sweden (NBIS) – Sweden
[8] Linköping University – Sweden
[9] Clinical Genetics – Sweden
[10] Dept. of Molecular Medicine and Surgery – Sweden
[11] Dept. of Medical Sciences – Sweden
[12] Uppsala University – Sweden
[13] Dept. of Biochemistry and Biophysics – Sweden
[14] Stockholm University – Sweden
[15] Dept. of Cell and Molecular Biology – Sweden
[16] School of Biotechnology, Division of Gene Technology – Sweden
[17] Royal Institute of Technology (KTH) – Sweden

We present Sarek, a portable Open Source pipeline to resolve germline and somatic variants from WGS data: it is written in Nextflow, a domain-specific language for workflow building. It processes normal samples or normal/tumor pairs (with the option to include matched relapses).

Sarek is based on GATK best practices to prepare short-read data, which is done in parallel for a tumor/normal pair sample. After these preprocessing steps several variant callers scan the resulting BAM files: Manta for structural variants; Strelka and GATK HaplotypeCaller for germline variants; Freebayes, MuTect2 and Strelka for somatic variants; ASCAT and Control-FREEC to estimate sample heterogeneity, ploidy and CNVs. At the end of the analysis the resulting VCF files can be annotated by SNPEff and/or VEP to facilitate further downstream processing. Our ongoing effort focuses in filtering and prioritizing the annotated variants.

Sarek is based on Docker and Singularity containers, enabling version tracking, reproducibility and handling sensitive data. It is designed with flexible environments in mind, like running on a local fat node, a HTC cluster or in a cloud environment like AWS. The workflow is modular and capable of accommodating further variant callers. Besides variant calls, the workflow

[*] Speaker

[†] Corresponding author: maxime.garcia@scilifelab.se

provides quality controls presented by MultiQC. Checkpoints allow the software to be started from FastQ, BAM or VCF. Besides WGS data, it is capable to process inputs from WES or gene panels.

The pipeline currently uses GRCh37 or GRCh38 as a reference genome, it is also possible to add custom genomes. It has been successfully used to analyze more than two hundred WGS samples sent to National Genomics Infrastructure (Science for Life Laboratory) from different users. The MIT licensed Open Source code can be downloaded from GitHub.

The authors thank the Swedish Childhood Cancer Foundation for the funding of Barntum´orbanken. We would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure, NGI, and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure.

Keywords: Somatic, Germline, Variations, Cancer, Workflow, Container, Reproducibility, Nextflow, Pipeline, Singularity, Docker, Genomics

Proteomic and phosphoproteomic analysis of medulloblastoma reveals distinct activated pathways between subgroups

Loredana Martignetti * [1], Laurence Calzone , Arnau Montagud , Stéphane Liva , Alexandre Sta , Patrick Poullet , Emmanuel Barillot

[1] Cancer et génôme: Bioinformatique, biostatistiques et épidémiologie d'un système complexe – Inserm : U900, Institut Curie, MINES ParisTech - École nationale supérieure des mines de Paris, PSL Research University, Paris – 26 rue d'Ulm - 75248 Paris cedex 05, France

In modern oncology, cell-molecular heterogeneity of human tumors is believed to be a major cause of drug resistance and subsequent disease recurrence. Therefore, the goal of precision medicine is to dissect complex molecular profile of each patient's tumor to understand mechanisms underlying disease aggressiveness and drug resistance. Recent large-scale genomic experiments are providing more detailed molecular characterizations of tumors, bringing the possibility of a more accurate stratification of tumor subtypes. To date, mainly owing to the maturity and availability of high throughput DNA- and RNA- based techniques, molecular classifications of tumors primarily focus on genomics and transcriptomics. Protein-level measurements are underutilized due to technical major hurdles including reproducibility. Recent advances in mass spectrometry (MS) have enabled extensive analysis of cancer proteomes. In this study, we employed quantitative proteomics to profile protein expression across 40 patient-derived samples of medulloblastoma (MB), which is the most common malignant brain tumor of childhood. During many years, MB was thought of as a single disease and thus, clinicians have been applying conventional therapy to all patients. However, advances have changed this single disease view and revealed a more complex reality. Tumor characterization from pathologists as well as recent extensive analysis using transcriptome and methylome profiling clearly demonstrated that MB comprises four distinct subgroups with unique molecular characteristics and patient outcome (Northcott et al., 2012). Nevertheless, a better understanding of specific signaling alterations and reliable diagnostic markers for the different MB subgroups is still needed.

Here, an integrative analysis was carried out including methylation, transcriptomics, small RNAs profiling as well as cutting-edge deep proteomics and phosphoproteomics measurements of MB to decipher signaling pathways and molecular mechanisms underlying different tumor subtypes. Using the recent approach, super-SILAC followed by MS analysis _˜6000/7000 proteins, _˜9.000/11.000 unique phospho-peptides and _˜2000/3000 unique phospho Tyr-peptides from MB samples have been accurately quantified.

We used network models that include multiple levels of information to integrate genome-wide molecular profiles for capturing the heterogeneity of observed phenotypes, leading to the identification of homogeneous subgroups of patients with similar disease outcome and to the identification of specific molecular features that characterize different subgroups. Regulatory gene networks have been previously used to investigate tumor heterogeneity and the development of

*Speaker

cancer cell plasticity. The approach applied here is distinct in that it uses the network of individuals as a basis for molecular data integration. The data produced by -omics experiments were used to construct a network of individuals. A patient-similarity network has been constructed to study the tumor heterogeneity based on multi-level molecular profiles (Wang B et al, 2014). A similarity measure for each pair of individuals was used to construct a patient-by-patient similarity matrix from each molecular data type. This matrix is equivalent to a similarity graph where nodes are individuals and the weighted edges represent pairwise similarities. Then, to create a comprehensive view of the patient subgroups combining the different molecular levels, we applied a network fusion method (Wang B et al, 2014) to integrate networks from different molecular profiles into a final consensus network. The peculiarity of this consensus network approach resides in the loss of weak similarity edges that helps to reducing noise while strong similarities in one or more networks are added to the final one. In addition, low-weight edges supported by multiple networks can be maintained. Therefore, the consensus network is able to detect both common and complementary information from different data levels. To extract relevant biological information from the structure of the constructed network, we studied its partition into clusters. This graph analysis offered insights into how informative each molecular level is to the observed disease variability. The detection and characterization of cluster structure in networks, meaning the appearance of densely connected groups of individuals, with only sparser connections between groups (communities), could be associated to relevant biological information. We investigated whether the communities within the similarity graph could correspond to subgroups of patients with similar responses to treatment and outcome of the disease. This integrative analysis clearly separated four distinct clusters aligned with the four subgroups of basic MB. We investigated molecular characteristics of these subgroups, in terms of active/inactive signaling pathways and transcritpional programs. Our results uncover deregulations in some signaling pathways specific to poor outcome subgroups, not previously apparent in transcriptomic comparisons. Overall, our integrative proteogenomic approach identifies a previously unknown oncogenic pathway and potential therapeutic vulnerability in the most common medulloblastoma subgroup.

References

Northcott, Paul A., et al. "The clinical implications of medulloblastoma subgroups." Nature reviews Neurology 8.6 (2012): 340.
Wang, B., et al. "Similarity network fusion for aggregating data types on a genomic scale." Nature methods 11.3 (2014): 333-337.

Keywords: Quantitative proteomics, multi level network, cancer, personalized medicine

From individual genetic variations towards haplotype: GEMPROT, a new way of reading VCF files

Tania Cuppens *† 1, Thomas Ludwig 1, Pascal Trouvé 1, Emmanuelle Genin 1

1 UMR1078 "Génétique, Génomique Fonctionnelle et Biotechnologies", INSERM, EFS, Université de Brest, IBSAM, CHU de Brest – INSERM U1078 – France

The development of next-generation sequencing (NGS) in the last decade has boosted genetic research and allowed a much finer characterization of the human genome. However, this finer characterization also comes with a price: we have to deal with load of information and finding ways to find the interesting "needles in a stack of needles". Indeed, we all have hundreds of thousands of genetic variants in our genome compared to the human reference genome; most of them are neutral and have no impact on our health. When sequencing a patient's genome to find the causes of the disease, it is necessary to filter these neutral variants to focus on those that may be involved in the disease. To do this, we usually consider the variants present in each gene one by one taking no account of the combinations of variants carried by each individual. This could be problematic as it is possible that genetic variants that have no effect when taken individually become deleterious when present in "cis" on the same haplotype. This could typically be the case for variants that are located in a same functional domain of the protein or variants located in different domains that interact and get in contact in the folded protein. The problem however is that it is not so easy to visualize these combination of variants when analyzing sequencing data as the vcf-files display the variants one per line.
To allow such haplotype analysis, we have developed GEMPROT, a bioinformatic tool to visualize the variations of the protein sequence induced by the genetic variations present in an individual. GEMPROT makes it possible to highlight combinations of variations that may affect the same functional domain of the protein. There was no tool so far to address this issue and we believe, this tool could be really helpful in sequence analysis.

GEMPROT is freely available and can be downloaded at https://github.com/TaniaCuppens/GEMPROT. It requires as input a phased VCF file, containing all the variants carried by the individual on its two haplotypes. The user then provides the name of the gene of interest and the exons positions of the gene are retrieved using NCBI data, based on the CCDS (Consensus Coding Sequence) Project (Pruitt et al., 2009). The nucleotide sequence of the transcript is then extracted from the reference genome using SAMtools (Li et al., 2009). This reference sequence is duplicated to represent the two haplotypes and modified when one or more mutations are present in the individual on one or the other haplotype.

*Speaker

†Corresponding author: tania.cuppens@inserm.fr

The two haplotypes are translated using a dedicated Perl script that reads through the entire nucleotide sequence taking into account all the mutations. Proceeding in this way rather than translating each mutation one by one as it is done by most sequence annotation tools avoid misinterpreting amino-acid changes when two mutations are found in cis in the same codon. Indeed, annotation tools usually consider the double mutations present on the same codon as two disjointed mutations leading to different amino acid changes and this could lead to translation errors that could impact the interpretation.

GEMPROT then displays, for each individual, the variations on each of the two protein sequences. These variations are mapped to the various known functional domains of the protein that are reconstructed using the pfam database. Modified amino acids are also associated with their physicochemical properties to provide a better interpretation of expected changes at the individual level. When the input vcf-file contains data on multiple individuals GEMPROT lists the different haplotypes found in the sample and their respective frequencies. For stratified populations, GEMPROT can also show the distribution of haplotypes across the different subgroups of individuals in order to evidence differences.

The different results are obtained in html format and a web version of GEMPROT is also available that allows to run the program on small datasets in a user-friendly environment (http://med-laennec.univ-brest.fr/GEMPROT/).
By offering a global visualization of the gene with the genetic mutations present, we believe GEMPROT could contribute to a better understanding of the impact of mutations combinations on the protein sequence and would allow to go beyond single locus analysis of sequencing data.

Keywords: Bioinformatics, Genetics, Protein sequence analysis, Visualization, Sequence analysis, Haplotypes

Bayesian Genome-Wide Association Study to discover novel lifespan-associated loci

Ninon Mounier *† [2,1], Paul Timmers [3], Kristi Läll [5,4], Krista Fischer [4], Zheng Ning [6], Xiao Feng [7], Andrew Bretherick [8], David Clark [3], Xia Shen [6,3], Tõnu Esko [4], James Wilson [8,3], Peter Joshi [3], Zoltán Kutalik [1,2]

[2] Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland – Switzerland

[1] Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital, Lausanne 1010, Switzerland – Switzerland

[3] Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, United Kingdom – United Kingdom

[5] Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia – Estonia

[4] Estonian Genome Center, University of Tartu, Tartu, Estonia – Estonia

[6] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden – Sweden

[7] tate Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Resources, Key Laboratory of Biodiversity Dynamics and Conservation of Guangdong Higher Education Institutes, School of Life Sciences, Sun Yat-sen University, Guangzhou – China

[8] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, United Kingdom – United Kingdom

Identification of genetic variants associated with a phenotype is the first step towards biomarker discovery. However, many Genome-Wide Association Studies (GWASs) are underpowered to detect such variants. Increasing the sample size also increases the GWAS power, but large datasets can be difficult to gather for certain traits, such as lifespan. Leveraging independent sources of information and include them as priors in a Bayesian analysis can improve GWAS studies without increasing sample size.
We developed a framework for informed GWAS (implemented in the R package bGWAS) that accounts for the prior information by comparing the observed Z-scores from a conventional GWAS to prior effects using Bayes Factors (BFs) to quantify evidence in favour of the prior (BF$>$ 1). Significance is assessed by calculating the probability of observing a value larger than the observed BF (P-value) given the prior distribution by decomposing the analytical form of the BFs and taking advantage of the fact that most SNPs have a zero prior effect estimate. For small BFs (i.e. insignificant P-values), an approximation is used to make the computation 10-times faster. This approach does not estimate posterior effects but allow a quick identification of variants for which prior and observed effect are consistent.

We applied this method to improve the power of a lifespan GWAS based on over 1 million parental lives. One way to derive prior effects is to combine summary statistics of GWASs for related traits with their causal effect on the trait of interest (using multivariate Mendelian Randomization). Our approach identified 12 traits significantly affecting lifespan, including BMI

*Speaker

†Corresponding author: mounier.ninon@unil.ch

(beta = –0.136, P = 1.0x10-27), smoking (beta = –0.442, P = 8.4x10-16), education level (beta = +0.216, P = 3.5x10-73), coronary artery diseases (beta = –0.239, P = 2.4x10-15), type 2 diabetes (beta = –0.090, P = 2.4x10-5) and insulin (beta = –0.172, P = 8.8x10-3). The prior effects derived from these risk factors lead to the identification of 10 new genome-wide significant variants in addition to the 25 identified by standard GWAS, notably in the long-suspected *ABO* gene (rs2519093 - increase of 2.6 months per C-allele, P = 5.5x10-10) and in *LPL* (rs1581675 - increase of 2 months per A-allele, P = 9.7x10-9). Most of the loci identified showed pleiotropic effects, such as a variant near *POM121C* (rs113160991 - increase of 2 months per G-allele, P = 2.2x10-9), which has not been significantly associated with any of the risk factors previously but might be affecting lifespan through moderate effects on insulin (Z-score = –4.52) and BMI (Z-score = –4.57).

So far, our priors have been built using information from summary statistics from other human GWASs, allowing us to only identify variants affecting lifespan through risk factors included in the prior but our method is adaptable and can be modified to use other types of prior information. As an example, it has been shown that the expression levels of specific genes can be used as a measure of biological age (revealing accelerated aging, hence shorter lifespan). Interestingly, we observed that variants regulating genes whose expression had been shown to vary with age are 2 times more likely (95% confidence interval: 1.5-2.9) to be associated with lifespan (at P< 5x10-5). This suggests that differential expression of age-related genes is not only a biomarker of aging, but some of them may directly influence lifespan. Capitalizing on this result, we are currently working on defining priors based on gene-expression data, using a cross-species approach to overcome the lack of human expression datasets coupled with lifespan, in order to identify variants associated with healthy aging.

Keywords: Genetics, GWAS, lifespan, aging, Bayes Factor, Bayesian, variant analysis

Aggregation of rare family-specific variants associated with Rheumatoid Arthritis

Maëva Veyssiere *† 1, Javier Perea 1, Laetitia Michou 2, Anne Boland 3, Vincent Meyer 3, Christophe Caloustian 3, Robert Olaso 3, Jean-François Deleuze 3, François Cornelis 4, Elisabeth Petit-Teixeira 1, Valérie Chaudru 1

1 GenHotel – Univ Evry, University of Paris Saclay – 91057, Evry, France
2 Division of Rheumatology, Department of Medicine, CHU de Québec-Université Laval – Québec, QC, Canada
3 Centre National de Recherche en Génomique Humaine – François Jacob Institute, CEA – Evry, France
4 GenHotel-Auvergne – Auvergne University, Genetic Department, CHU Clermont-Ferrand – Clermont-Ferrand, France, France

1. Introduction

Rheumatoid arthritis (RA) is one of the most frequent autoimmune disease, affecting 0.3 to 1% of the population worldwide. Since the discovery of the association of the HLA locus with RA, around 100 genetic factors were identified as disease susceptibility loci by Genome Wide Association Studies (GWASs) or meta-analysis of GWASs. However, the effect size of most of these genetic risk factors is too weak to explain the entire RA genetic component. Indeed, the proportion of heritability attributed to HLA-DRB1 shared-epitope alleles was estimated between 11% [1] and 37% [2]. The other loci, identified by GWAS outside the HLA locus, explain only an additional 5% of the RA heritability. The identification of new susceptibility loci would allow a better understanding of RA pathogenesis, and would help to develop new therapeutic targets or biomarkers useful for an earlier diagnosis of RA.

Several hypotheses have been proposed to explain this missing heritability, one of them is the ”common disease – rare variants” hypothesis. This paradigm states that multiple rare variants, each of them having a relatively high penetrance, could be involved in common diseases like RA. Rare variants, not identifed by GWASs which allow mainly to analyse common genetic variants, are detectable by Whole Exome Sequencing (WES). Nevertheless, despite easier access to sequencing, the discovery of rare variants is still challenging. Indeed, the power of association signal between a rare variant and a disease can vary according to the type of study and the chosen strategy (e.g: case-control study or family-based study, univariate or burden test).

Here we present a study based on multiplex families (at least 4 RA cases in the first degree relatives) designed to identify rare variants that could play a role in the development of RA.

2. A family-based study

To allow a better identification of rare variants and overcome the problem of population stratification encountered with cases-controls studies, we chose a family-study design. In this study, we had access to 16 French multiplex families with affected carrying at least one shared epitope

[*] Speaker

[†] Corresponding author: maeva.veyssiere@univ-evry.fr

allele. We selected 30 individuals as a discovery set and included all 110 available individuals in the validation set. The discovery set was composed of 19 RA cases and 11 unaffected relatives belonging to 9 of the previously described pedigrees. We used this set to perform WES and identify RA-susceptibility variants. The validation set, consisting in 50 RA affected and 60 unaffected individuals, was used to perform resequencing of the candidate variants and validate their association with RA.

3. Identification of variants in whole exome sequences

We captured the exons in the discovery set with Agilent SureSelect Human All Exon kit (V5) and sequenced them on an Illumina HiSeq2000 platform. Then, we mapped the reads to the human reference genome hg19 using BWA-MEM algorithm [3] and removed the duplicates with Picard toolkit [4].

Variants were called in the targeted regions, plus 150 bp up and downstream, by using Haplotype Caller (HC) algorithm from the GATK suite [5]. Finally, to obtain high quality data, we applied the following filters. First, we selected SNVs and small indels (maximum length of 50 bp) complying with the following rules: total read depth DP $\geq$ 12, mapping quality MQ $\geq$ 30, variant confidence QD $\geq$ 2, strand bias FS score $\leq$ 25 and call-rate $\geq$ 95%. Then, we kept DNA variations outside segmental duplications and repeated regions, such as described in RepeatMasker from UCSC.

To assess the remaining variants frequency in reference European population, we extracted the minor allele frequency (MAF) from public databases for populations with European ancestor origin. We used four datasets: the 1000 Genomes Project, the Exome Aggregation Consortium project, the Exome Sequencing project and the Complete Genomics project.

4. RA candidate variants selection

We focused our research of RA candidate variants in a pool of 21,532 rare high-quality variants, with a Minor Allele Frequency under 1%.

Under the statement that RA rare variants in multiplex families have a high penetrance, we selected 2,143 variants segregating in all affected individuals within one family. We evaluated the potential functional effects of such variants using CADD phred-like score [6] and SNPeff [7]. And, we picked the top 0.01% genome-wide most deleterious variants, according to CADD annotation, predicted with HIGH or MODERATE effect by SNPeff. By these successive filters, we obtained 73 heterozygous variants candidates for RA from 73 different genes, not previously associated with the disease.

5. Evaluation of the RA genetic association at the gene level

We evaluated the RA-association of these 73 genes by performing a burden test with pVAAST [8]. Here, we considered a dominant model of inheritance, given the heterozygous nature of the candidate variants, and a maximum disease prevalence equal to 0.01 [9]. pVAAST requiring controls outside the tested pedigree, we added to the discovery set 45 matching individuals (European ancestor origin and sequencing on an Illumina platform) extracted from the 1000 genome project. We estimated disease association p-values by allowing the algorithm to perform up-to 1E+06 permutation tests.

Thirty-five genes (48% of the candidates) showed a significant association under nominal p-value of 0.05. We observed that in those genes, the variants with the highest score were the candidates. So, to validate the leading effect of these variants, we performed pVAAST test by including first, only the candidate SNVs, and then, all variants except the candidates. In total more than 90% of the tested genes were not anymore significantly associated with RA without the candidate variant.

6. Candidate variants validation in extended pedigrees

We pursued the analysis on the 10 candidate variants with the strongest genetic association (*ie* with the smallest p-values). Hence, we re-sequenced them with a different method in the validation set. To this end, we used the Ion Personal Genome Machine (PGM) System to sequence the targeted variants (the primers were designed with the Ion AmpliSeqTM Designer). For samples in both discovery and validation set, we checked the genotype concordance with whole-exome data.
We obtained exploitable data for 8 variants out of 10. If all of them remained family-specific, we noticed an aggregation of candidate variants (75% of re-sequenced variants) in one of our pedigrees called family 3.
In this family, composed of 4 affected and 5 unaffected individuals, we observed 4 non-synonymous and 2 nonsense variants. Two variants completely segregated among all cases and three co-segregated among all affected and one unaffected. The candidate variant with the strongest genetic association (p-value = 0.0029) belonged to the group of variant with complete segregation and was not reported to date. It is a nonsense variant observed in a gene involved in the regulation of macroautophagy, which plays a key role in the pathogenesis of RA [10]. This variant introduces a premature stop codon at the beginning of the gene.

7. Conclusion

The primary objective of this study was to identify new RA candidate loci to improve our knowledge about this disease genetic component. We highlighted, through WES of RA multiplex pedigrees, the aggregation of multiple RA risk loci specific to one of the sampled families. We identified in particular one nonsense variant, never reported to date, within a gene regulating the process of macroautophagy. *In vitro* functional studies could help to characterize this variant pathogenicity and its impact on autophagy.
We should re-sequenced the candidate genes in a different dataset to validate their association with RA and identify new potential causal variants.

8. References

1. Van der Woude Diane, Houwing-Duistermaat Jeanine J., Toes René E. M., Huizinga Tom W. J., Thomson Wendy, Worthington Jane, et al. Quantitative heritability of anti–citrullinated protein antibody–positive and anti–citrullinated protein antibody–negative rheumatoid arthritis. Arthritis Rheum. 2009;60: 916–923. doi:10.1002/art.24385

2. Deighton CM, Walker DJ, Griffiths ID, Roberts DF. The contribution of HLA to rheumatoid arthritis. Clin Genet. 1989;36: 178–182.

3. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. 2013; Available: http://arxiv.org/abs/1303.3997

4. Picard Tools - By Broad Institute [Internet]. [cited 13 Nov 2017].
Available: http://broadinstitute.github.io/picard/

5. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al. 2013;11: 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

6. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46: 310–315. doi:10.1038/ng.2892

7. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). /01//;6: 80–92. doi:10.4161/fly.19695

8. Hu H, Roach JC, Coon H, Guthery SL, Voelkerding KV, Margraf RL, et al. A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. Nat Biotechnol. 2014;32: 663–669. doi:10.1038/nbt.2895

9. Kurkó J, Besenyei T, Laki J, Glant TT, Mikecz K, Szekanecz Z. Genetics of rheumatoid arthritis - A comprehensive review. Clin Rev Allergy Immunol. 2013;45: 170–179. doi:10.1007/s12016-012-8346-7

10. Dai Y, Hu S. Recent insights into the role of autophagy in the pathogenesis of rheumatoid arthritis. Rheumatology. 2016;55: 403–410. doi:10.1093/rheumatology/kev337

Keywords: Whole exome sequencing, Complex disease, Rare variants, Rheumatoid arthritis

Unravelling human preimplantation development by single-cell RNA-Seq: from experiment design to cell fate trajectories.

Dimitri Meistermann *† [1,2], Julie Firmin [1,3], Stéphanie Kilens [1], Sophie Loubersac [1], Valentin François–Campion [1], Betty Bretin [1], Thomas Fréour [3], Jérémie Bourdon [2], Laurent David [1]

[1] Centre de Recherche en Transplantation et Immunologie (CRTI) – Université de Nantes, Institut National de la Santé et de la Recherche Médicale : U1064 – CHU Nantes, 30 Bd Jean Monnet, 44093 Nantes Cedex 1, France

[2] Laboratoire des Sciences du Numérique de Nantes (LS2N) – Université de Nantes, Ecole Centrale de Nantes, Centre National de la Recherche Scientifique : UMR6004, IMT Atlantique Bretagne-Pays de la Loire – Université de Nantes – faculté des Sciences et Techniques (FST)2 Chemin de la HoussinièreBP 92208, 44322 Nantes Cedex 3, France

[3] CHU de Nantes, Service de Biologie de la Reproduction – CHU Nantes – France

Single-cell based high-throughput sequencing technologies have led understanding of cell biology to a new perspective by investigating systematically the fundamental unit of this field. Thus, domains with highly-heterogeneous material have beneficiated of these advances, such as immunology (Papalexi and Satija, 2017) and oncology (Liang and Fu, 2017). This understanding of cell heterogeneity has also brought a new way to study tissues complexity, allowing to begin project such as Human Cell Atlas, with the aim to mapping all human cell types (Regev et al., 2017). Single-cell biology has also raised considerable interests concerning its ability to study differentiation processes, especially in early development biology, while the human embryo just before its implantation in utero has less than few hundreds of cells and is composed by several lineages.
The major goal of our team is to understand and quantify human preimplantation development, from fertilization to implantation in the uterus, and to predict its outcome. Specifically, we aim at deciphering the molecular mechanism driving cell fate during this first step of our existence. Understanding human preimplantation is therefore critical to improve assisted reproductive technologies (ART) and broaden the use of human pluripotent stem cells in regenerative medicine. It is during this timeframe that embryonic cells make their first choice of cellular fate, moving from one totipotent cell in the zygote to an embryo stratified by three cell types in the mature blastocyst. Moreover, the main objective of in vitro fertilization is to support the development of the zygote into a blastocyst, before transfer in infertile patients. To discover how early cell fate specification is regulated, our team develops several strategies to model embryos. One of the major strategy we used is single-cell RNA-Seq. In this oral communication I will present the path we took, from our question to the analysis of cell fate trajectories.

The first step of single-cell RNA-Seq is designing the experiment. 31 embryos from the post

[*] Speaker

[†] Corresponding author: dimitri.meistermann@univ-nantes.fr

fertilization day 3 to 6 were gathered in the IVF clinic of Nantes. We obtained a total of 388 cells. Each embryo was filmed during its development by a time-lapse machine and annotated by clinicians. This gave us precious information on the morpho-kinetics aspect of preimplantation development and completes the transcriptome study. Once the sequencing in done, the first analysis phase consists in quality control and mapping of reads. These are the classic first steps of RNA-Seq analysis. At this point, the analysis workflow begins to bifurcate with bulk RNA-Seq. Hence we used a single-cell specific workflow based on the SCRAN package (Lun et al.) for normalizing data.

Recently, a large dataset of 1529 cells from human preimplantation embryo was published (Petropoulos et al., 2016), with the same sequencing strategy as our: SMART-Seq2. This allowed us to merge both dataset. We were also able to reannotate Petropoulos' dataset based on similarity on theirs cells to ours. Indeed, we have additional annotations in our data, such as morpho-kinetic data and side of the cell on the embryo before sequencing.

With the reads count table of each cell, we decided to carry out two different complementary analysis. The first is a gene-based analysis, Weighted Gene Correlation Analysis (WGCNA). Briefly, WGCNA split the genes library into modules from on a co-expression matrix. In addition of the module attribution of genes, WGCNA give a contribution matrix of each sample in each module. With this tool, we determined specific gene networks associated with subpopulations of cells, embryo lineage or embryo stage. Thus, WGCNA is a particularly suitable tool for single-cell analysis, giving coherent results from noisy and heterogenous data. The second analysis is the pseudotime estimation through Monocle2 (Qiu et al., 2017). Monocle2 needs a subset of genes to work properly called "ordering genes". Ordering genes must be highly variable for distinguishing cells and sorting them into different branches. We selected genes that were over-dispersed among datasets as ordering genes. We tried several thresholds for picking up ordering genes. To help us in the decision of the best threshold, we calculated a benchmark score based on the concordance between Monocle2 results and our cell annotations. After the choosing of ordering genes, Monocle2 makes a reversed graph embedding projection, a nonlinear dimensionality reduction algorithm. This type of procedure allows to retrieve processes that generated data. In our case, this results in retrieving cell fate trajectories. Finally, Monocle2 give a value to each cell to quantify its distance to the root. This value is called "pseudotime". Thus, the pseudotime delta gives a measure of transcriptomic changes intensity.

With each face of this project, we were able to integrate each aspect of the data. As major result we saw that intensity of transcriptomic changes were no the same for each preimplantation cell fate. We also gave a precise chronology of the human preimplantation development events. Moreover, we combined WGCNA and pseudotime analysis to visualize genes waves apparition through the pseudotime and defining potential cell fates precursor genes. Some of the best gene candidate as expression lineage marker or precursor were validate with immunofluorescence on embryo. With the SMART-Seq2 technology, we were also able to track changes in the gene allelism through the pseudotime and examine phenomenon as X chromosome inactivation.

This work is done simultaneously with another project of the team that concerns cell reprogramming. In a preceding work (Kilens et al., 2018) we showed that induced na´ive pluripotent stem cells are a good in vitro model for the epiblast, the blastocyst lineage that give rise to the individual. The next step I to retrieve an in vitro model for the trophectoderm, the second lineage of the blastocyst, that give rise to the placenta. This goal is considerably facilitated by the single-cell analysis of blastocysts.

Finally, we will release our analysis with a user interface based on d3.js for sharing our data with the community, and for instance give the possibility for each team to observe its gene of interest on our model. We hope that our work will contribute significantly to our understanding of preimplantation development and will open new avenues of research in the fields of ART and regenerative medicine fields.

References

Kilens, S., Meistermann, D., Moreno, D., Chariau, C., Gaignerie, A., Reignier, A., Lelièvre, Y., Casanova, M., Vallot, C., Nedellec, S., et al. (2018). Parallel derivation of isogenic human primed and naive induced pluripotent stem cells. Nature Communications *9*.

Liang, S.-B., and Fu, L.-W. (2017). Application of single-cell technology in cancer research. Biotechnology Advances *35*, 443–449.

Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. https://master.bioconductor.org/packages/release/workflows/v
1-reads.html.

Papalexi, E., and Satija, R. (2017). Single-cell RNA sequencing to explore immune cell heterogeneity. Nature Reviews Immunology *18*, 35–45.

Petropoulos, S., Edsg´ard, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell *165*, 1012–1026.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nature Methods.

Regev, A., Teichmann, S., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The Human Cell Atlas.

Keywords: single cell, Preimplantation development, pseudotime, RNA Seq, pluripotency, human, integrative analysis, cell fate

Fine-scale genetic population structure in western France

Christian Dina *† [1], Joanna Giemza [2], Matilde Karakachoff [3], Karen Rouault [4], Floriane Simonet [3], Claude Férec , Hervé Le Marec [3], Stéphanie Chatel [3], Jean-Jacques Schott [3], Emmanuelle Genin [4], Richard Redon [3], Christian Dina‡ [3]

[1] institut du thorax (U1087) – Institut National de la Santé et de la Recherche Médicale - INSERM : UMR1087, Centre national de la recherche scientifique - CNRS (France) : UMR6291, Université Nantes Angers Le Mans, CHU Nantes – France

[2] institut du thorax (u1087) – Institut National de la Santé et de la Recherche Médicale - INSERM : UMR1087, Centre National de la Recherche Scientifique - CNRS : UMR6291 – France

[3] institut du thorax (U1087) – Institut National de la Santé et de la Recherche Médicale - INSERM : UMR1087, Centre national de la recherche scientifique - CNRS (France) : UMR6291 – France

[4] UMR 1078 – Institut National de la Santé et de la Recherche Médicale - INSERM : UMR1078 – France

Introduction:
Characterizing the genetic structure of human populations provides insight into demographical history and informs research on disease association studies, especially on rare recent variants which tend to be geographically clustered.

Our team studies genetic structure in North-Western France. We have previously shown genetic proximity between Bretons and Irish in Karakachoff et al as well as high correlation between geography and genetics at the fine regional scale and higher differentiation between Britanny and neighboring regions. Moreover, we also reported preliminary results of possible higher inbreeding in the western-most part of this region.

In the present study, we examine the fine-scale genetic structure of Brittany, Anjou, Poitou and Maine in western France in a new dataset. Our area od study is a large peninsula positioned in the northwest of France and delimited by the English Channel to the north, Atlantic Ocean to the west and the Bay of Biscay to the south; the whole Region covers approximately 60 000km2 extending approximately 470 km from west to east.

We then applied standard population genetic methods on a representative set of 3200 individuals, to identify the nature of genetic structure as well as exploring the demographic history and possible consequences on medical research in this 7M inhabitants region.

Methods:

We genotyped 2,500 individuals whose 4 grandparents were born within a small distance in west-

[*] Speaker

[†] Corresponding author: christian.dina@univ-nantes.fr

[‡] Corresponding author: christian.dina@univ-nantes.fr

ern France on AxiomTM Precision Medicine Research Array (Affymetrix - Thermo Fisher Scientific). array plates. The individuals were either analyzed based on the centroid of birth place of the 4 grand-parents, or at the level of two administrative structures, the "département" and the fine level of "arrondissement". On average, we have 4 "arrondissements" for a "département". While these are administrative structures, they tend to reflect some historical information – for instance the modern Maine-et-Loire is covering the historical duchy of Anjou.

Principal Component Analysis (PCA) and Fst between departments

Principal Components Analysis was carried out using the smartpca software from the EIGENSOFT package version 6.0.1 (Price, Patterson et al. 2006).

To evaluate the geographic relevance of PCs, we tested for significance of association between latitude, longitude of each department and PCs coordinates ('cor.test' function in R) using a Spearman's rank correlation coefficient.

Level of genetic differentiation, Fst values between administrative units were obtained with smartpca software. The road distances, to evaluate distance between villages, were obtained thanks to 'map.dist' R package.

Chromopainter/FineSTRUCTURE analysis

We applied a recently developed method for investigating fine-scale population structure of French, Chromopainter version 2 and FineSTRUCTURE version 2.0.7 (Lawson, Hellenthal et al. 2012).

The Chromopainter algorithm reconstructs an individual's chromosomes as a series of genomic fragments from potential donor individuals in the data set. In practice, we 'painted' the chromosome of every individual (receiver) in a data set using the haplotypes of all other individuals (donors).Chromopainter requires an initial phasing step. The datasets were phased using SHAPEIT v2.r790 (Delaneau, Marchini et al. 2012) (Delaneau, Zagury et al. 2013) and the genetic map build 37 provided with that software.

FineSTRUCTURE was used to reconstruct a tree that infers population relationships and similarities between individuals, using the co-ancestry matrix generated with Chromopainter. We were therefore able to identify clusters of genetically close individuals.

Use of haplotype information is expected to provide deeper insight into the fine scale structure of a population.

Identity by Descent (IBD):

We used RefinedIBD, for homozygosity (HBD) and identity by descent (IBD)-based analyses. The two measures also report genetic distance either within (HBD) or between (IBD) individuals.

We will report average HBD length per-individual within "départements" and "arrondissements". Similarly, we are reporting the levels of identity by descent between individuals from the various administrative units.

IBD-estimated population size

IBDseq and IBDNe were used to infer effective population size and its growth rates. IBDseq detects segments of IBD while IBDne estimates the historical effective population size of a homogenous population for each of 150 generation backwards. The IBDNe software calculates the effective population size of a given population over past generations by modeling the distribution of IBD tract lengths present in the contemporary population.

Results:

Principal Components are highly correlated with geographical coordinates of grandparents' birthplaces (p-value $<$ 2e-16). Visualisation of single principal components' values (from PC1 to PC5) on the map reveals patterns of local genetic structure. Loire River and its tributaries, "Erdre" and "Sèvre Nantaise", seem to be at the limits of observed genetic subpopulations. Moreover, we also observe visually, possible genetic barrier at the level of what could have been historical border of Duchy of Britanny, where no clear geographical limit can be identified.

Analysis based on haplotype structure, using Chromopainter, confirmed this fine-scale structure, providing even higher correlation between geography and genetics.

The initial division happens between North and South of Loire, whereas the second division could also be interpreted as reflecting the political or linguistic border of Britanny.

Analysis based on haplotype structure, using Chromopainter, confirmed this fine-scale structure, providing even higher correlation between geography and genetics. The cluster obtained from this analysis are identifying even more detailed regions which need careful interpretation in terms of history and geography. 79 clusters with more than ten individuals – on average (79/9 = almost) 9 subgroups per department. Clusters are localized and almost non-overlapping.

Visualization of average length of runs of homozygosity on the departmental level allowed us to observe a pattern of gradual increase towards the end of the Brittany. Finer scale of arrondissement allows to notice local phenomena, in particular near Brest, St. Malo and Cholet.

The IBD analysis highlighted higher level of IBD sharing in the arrondissements of Brittany. In parallel, there is correlation between IBD segment length between individuals and geographical distance. The IBD resemblance decreases for individuals outside of Britanny.

Using IBD sharing (IBDNe), we were able to identify the recent exponential growth in population size already seen in European populations. Moreover, we also show that identified subgroups may have followed different trajectories of effective population size evolution in the last 25 generations.

This results hint that three subpopulations had different demographical history, in particular between 10-21 generations ago, the model indicates growth, plateau and bottleneck for Bretons, the population on the south of Loire River and the population extending to North-East, respectively. However, the signal should be interpreted with caution, as it might have been confounded by admixture or population structure.

Conclusion

We here report existence of a fine-scale structure across western France, with evidence of distinct demographic histories between subpopulations. To our knowledge this is among the first observations of a coincidence of geographical and historical limits with genetic barriers at such a small geographical level in Europe. The importance of both the Loire River and also of historical or linguistic borders as a possible genetic barriers in one European population is also a key observation.

These results support the need for a genetically matched panel of controls from France, to avoid confounding effects of fine-scale population structure.

From these observations, we conclude that the genetic structure of this region is shaped both by isolation by distance and existence of genetic barriers.
Further study of demographic models will yield not only insight on population history, but also provide a null model for tests of selection.

Keywords: population genetics Britanny history

Exploring Chemical Space Using ChemMaps.com

Alexandre Borrel *† [1], Nicole Kleinstreuer [2], Denis Fourches [3]

[1] Division of Intramural Research/Biostatistics and Computational Biology Branch, NIEHS (NIH/NIEHS) – Research Triangle Park, NC, United States
[2] Division of Intramural Research/Biostatistics and Computational Biology Branch, NIEHS, RTP, NC, USA – United States
[3] Department of Chemistry Bioinformatics Research Center North Carolina State University – United States

The need for navigating the chemical space has become more important due to the increasing size and diversity of chemical biological databases (*e.g.,* Chemspider, DrugBank, ChEMBL, Toxcast). To do so, modelers typically rely on projection techniques applied to series of quantitative molecular descriptors directly computed from two-dimensional chemical structures. However, the multiple cheminformatics steps required to compute and visualize a chemical space are technical, necessitate coding skills, and thus represent a real obstacle for non-specialists. Inspired by the popular Google Maps application, we developed the *ChemMaps.com* webserver to easily navigate chemical spaces. The first version of ChemMaps was developed to browse and visualize the space of 2,000 FDA-approved drugs and over 6,000 drug candidates. Each compound was initially characterized using a large set of molecular descriptors including 1D-2D RDKIT descriptors and 3D PADEL descriptors (238 1D-2D and 44 3D after removing correlated descriptors). Principal Component Analysis was used to project compounds in three-dimensional space, where compounds' coordinates in the first two dimensions were calculated using 1D-2D descriptors, and the third dimension (Z axis) was determined using 3D descriptors only. To optimize the representation of the space and the interactive, user-friendly navigation experience, we developed the *ChemMaps.com* webserver using modern 3D-optimized web technologies such as HTML5, JavaScript, and CGI. The chemical coverage is now being expanded to include environmental chemical space based on the U.S. EPA TSCA inventory, as well as toxicological categorizations based on curated animal study data and predictive high-throughput screening signatures. Users accessing *ChemMaps.com* can immediately explore the entire compound library using a responsive, mouse-based, easy-to-use navigation tool. Since all information and coordinates are pre-computed, the browsing is instantaneous and does not require computational skills. Similar to searching Google Maps for a specific address, users can search the ChemMaps via a dedicated search bar (*e.g.,* name, indications, pharmacological class, toxicity values) and visualize the space with options to zoom in on chemical "neighborhoods". Additional browsing, searching, and exporting options are underway, including tools to support read-across and chemical risk assessment.

*Speaker

†Corresponding author: a.borrel@gmail.com

Keywords: Chemical space, webserver, toxicology, data visualization

Interpretation of mass spectrometry-based metaproteomics: how much can we trust the MetaHIT 9.9 catalog?

Ariane Bassignani *† [1,2,3,4], Magali Berland‡ [1], Sandra Plancade [3], Catherine Juste [2], Olivier Langella [4]

[1] MetaGenoPolis (MGP) – Institut National de la Recherche Agronomique – Centre de Jouy-en-Josas Domaine de Vilvert F78352 JOUY-EN-JOSAS Cedex, France

[2] MICrobiologie de l'ALImentation au Service de la Santé humaine (MICALIS) – Institut National de la Recherche Agronomique : UMR1319, AgroParisTech – F-78350 JOUY-EN-JOSAS, France

[3] Mathématiques et Informatique Appliquées du Génome à l'Environnement (MAIAGE) – Institut National de la Recherche Agronomique – Centre de Jouy-en-Josas Domaine de Vilvert F78352 JOUY-EN-JOSAS Cedex, France

[4] PAPPSO, GQE - Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay – Institut national de la recherche agronomique (INRA) – 91190 Gif-sur-Yvette, France

The integrated gene catalog, comprising 9.9 millions of genes ('MetaHIT 9.9'), is a powerful reference tool to match metaproteomic mass spectra to genes of the human gut microbiota. However, given the individual specificities of the gut microbiota, one could fear that the MetaHIT 9.9 catalog would not cover the whole diversity, leading to a low level of mass spectra interpretation. In this work, we compared the performance of matching mass spectra against MetaHIT 9.9 or against individual metagenomic catalogs.

Using Ion Proton sequencers, we sequenced the gut microbiota of 236 individuals recruited under the FP7 MetaCardis framework. The reads were assembled to build specific gene catalogs for each sample, which were translated to generate 236 individual protein catalogs. Simultaneously, the gut microbiota of the individuals were extracted, and the cytosolic metaproteomes were analyzed on an Orbitrap Fusion Lumos Tribrid mass spectrometer. Identification of peptides and proteins was performed with X!TandemPipeline and compared using both the specific catalogs and the MetaHIT 9.9 catalog.

Compared to MetaHIT 9.9, the individual catalogs led to identify less peptides, but a close number of proteins. This means that the length of protein coverage was lower when using the individual catalogs. Interestingly some peptides were only identified with the individual catalogs, showing that they are useful to identify individual-specific peptides. As specific peptides can represent precious markers of various human disease phenotypes, this work lays the foundations for new biomarkers hunt strategies in personalized medicine.

Keywords: Gut microbiota, mass spectrometry, metaproteomic

*Speaker

†Corresponding author: ariane.bassignani@gmail.com

‡Corresponding author: magali.berland@inra.fr

Semantic lab. data interoperability: Encoding microbiology tests results using international biomedical onto-terminologies

Maël Le Gall * [1,2], René Vachon† [2], Xavier Gansel‡ [2]

[1] Université de Rouen Normandie (UNIROUEN) – Normandie University, UNIROUEN – 1, rue Thomas Becket 76821 Mont-Saint-Aignan Cedex, France
[2] bioMérieux - Lyon – BIOMERIEUX – France

To improve the patient care, health autorities want to make medical information sharing more fluid in the way to optimize treatment chain. In France, this thinning is through Electronic Health Record systems as DPI (i.e Dossier Patient Informatisé) or DMP (i.e. Dossier Médical Partagé).
Each IVD (i.e. In Vitro Diagnostic) systems transmits test descriptions (Observations[1]) and test results (Observation Values[1]) to the Laboratory Information Systems. The use of distinct vocabularies in EHR poses interoperability challenges. To pass through BIOMÉRIEUX proposes to encode their observation values with standard biomedical onto-terminologies such as LOINC®[2] (with LOINC® Answers) and SNOMED-CT®[3]. Usage of those terminologies would streamline injection of standardly described results in EHR systems.

This study aims to, from an IVD manufacturer perspective, (i) demonstrate if LOINC® Answers and SNOMED-CT® are capable of encoding a selected menu of microbiology IVD observation results, (ii) identify the technical road blockers for a global adoption of both terminologies and finally (iii) propose ways to circumvent those blockers.

BIOMÉRIEUX ordinal (i.e. pos., neg., detected, ...) observations results values for BACT/ALERT®, VITEK® 2, VIDAS®, ETEST® and BIONEXIA® (BIOMÉRIEUX system for micro-organisms identification, Antibiotic susceptibility testing and immunology testing) were encoded with LOINC® Answers and SNOMED-CT®. Simultaneously nominal (i.e. micro-organisms names) observation results values were mapped to SNOMED CT®. Both observations values were mapped using a combination of an internal string matching algorithm and manual curation of the hits involving BIOMÉRIEUX expert biologists and taxonomy references such as LPSN, DSMZ, Catalogue of Life, ICTV, etc.

We were able to map to SNOMED-CT® 90,5% (1335) of the 1476 nominal observation results values. Those nominal observation results values mapping cover 97,6% (1303) of consensual taxa (100% - 82 of genus; 97% - 1141 of species and 98% - 59 subspecies) and far less for non-consensual taxa such as 'x OR y' (23% - 11), variants (29% - 14) and groups (19% - 7). We also observed wide differences in code coverage between bacteria, yeasts and filamentous fungus.

*Speaker

†Corresponding author: rene.vachon@biomerieux.com

‡Corresponding author: xavier.gansel@biomerieux.com

Regarding the 15 unique ordinal observation results values, we mapped 66% (10/15) with LOINC® Answers and 60% (9/15) with SNOMED-CT®. Each ordinal observation results value is used for more than one observation.

This study shows that concerning nominal observation results values SNOMED-CT® covers a large majority of our selected menu results. Regarding ordinal observation results values we used LOINC® Answers and SNOMED-CT® which gave similar results, but it seems that the community is more oriented toward SNOMED-CT®. Most of missing codes were for (i) complex ordinal Observation values, (ii) non-consensual taxa and (iii) filamentous fungi and yeast. Looking across our results and the orientation of health IT community, it appears that SNOMED-CT® is the solution to encode microbiology tests results. Solving the above mentioned road blockers will require a dedicated effort synchronized with SNOMED International.

References:

Understanding Observations and Observation Values ; https://www.healthit.gov/isa/node/1096

What is LOINC® ; https://loinc.org/
What is SNOMED-CT ; https://www.snomed.org/snomed-ct

Keywords: SNOMED, CT, LOINC answers, interoperability, EHR

AnnotSV: An integrated tool for Structural Variations annotation

Véronique Geoffroy * [1], Yvan Herenger [2], Arnaud Kress [3], Corinne Stoetzel [1], Amélie Piton [4,5], Hélène Dollfus [1,6], Jean Muller [1,4]

[1] Laboratoire de Génétique médicale
(UMR$_S INSERMU$1112) – $-InstitutNationaldelaSantéetdelaRechercheMédicale - INSERM -$
$-IGMA, FacultédeMédecineFMTS, UniversitédeStrasbourg, Strasbourg, France$
[2] Service de Génétique Médicale – CHU Tours – Tours, France
[3] ICUBE UMR 7357, Complex Systems and Translational Bioinformatics (CSTB) – Université de Strasbourg - CNRS – Strasbourg, France
[4] Laboratoires de Diagnostic Génétique – Les Hôpitaux Universitaires de Strasbourg (HUS) – Institut de Génétique Médicale d'Alsace (IGMA), Hôpitaux Universitaires de Strasbourg, Strasbourg Cedex, France
[5] Institut de Génétique et de Biologie Moleculaire et Cellulaire (IGBMC) – INSERM U964, CNRS UMR7104, Université de Strasbourg – Illkirch, France
[6] Centre de référence pour les Affections Rares en Génétique Ophtalmologique (CARGO) – Filière SENSGENE, Hôpitaux Universitaires de Strasbourg – Strasbourg, France

Structural Variations (SV) are a major source of variability in the human genome that shaped its actual structure during evolution. Moreover, many human diseases are caused by SV, highlighting the need to accurately detect those genomic events but also to annotate them and assist their biological interpretation in the personalized medicine era.
Therefore, we developed AnnotSV that compiles functionally, regulatory and clinically relevant information and aims at providing annotations useful to i) interpret SV potential pathogenicity and ii) filter out SV potential false positive. In particular, AnnotSV reports heterozygous and homozygous counts of single nucleotide variations and small insertions/deletions called within each SV for the analyzed patients, this genomic information being extremely useful to support or question the existence of an SV. We also report the computed allelic frequency relative to overlapping variants from DGV (MacDonald, et al., 2014), that is especially powerful to filter out common SV.

To delineate the strength of AnnotSV, we annotated the 4,751 SV from one sample of the 1000 Genomes Project, integrating the sample information of 4 million of SNV/indel, in less than 60 seconds.
AnnotSV is implemented in Tcl and runs in command line on all platforms. The source code is available under the GNU GPL license. Source code, README and Supplementary data are available at http://lbgi.fr/AnnotSV/. Moreover, in order to provide a ready to start installation of AnnotSV, each annotation source (that do not require a commercial license) is already provided with the AnnotSV sources.

[*] Speaker

Keywords: Medical genomics, structural variations, annotation

A new protocol for sequencing and analysis of virus genomes in clinical context with amplicon-seq data.

Florence Maurier * [1,2], Delphine Beury [1,2], Anne Goffard [2], David Hot [1,2], Ségolène Caboche [1,2]

[1] PEGASE-Biosciences – Institut Pasteur de Lille – France
[2] Centre d'Infection et d'Immunité de Lille (CIIL) - U1019 - UMR 8204 (CIIL) – Institut Pasteur de Lille, Institut National de la Santé et de la Recherche Médicale, IFR142 : U1019, Université de Lille, Centre National de la Recherche Scientifique, UMR 8204 – 1 Rue du Professeur Calmette - Lille Cedex - 59019 - BP 245, France

High-throughput sequencing provides good opportunity for the large scale sequencing of virus genomes. However, despite the relative small size of virus genomes, their sequencing often remains difficult. Indeed, the small amount of virus RNA, compare to the host nucleic acid, requires specific treatments to enrich or amplify the virus genomes. In addition, viruses are present as a population of sequences rather than a unique genome sequence. This population is more or less variable depending on the intrinsic mutation rate of the virus complicating the final assembly of the genome.
Here we present an innovative protocol developed to sequence and analyze virus whole genomes for a routine use in a clinical context. This protocol was applied to RNA virus genomes extracting from 13 samples of patients infected with HCoV-OC43 coronaviruses isolated at the University Hospital of Lille. Belonging to the family of Coronaviridae, to the beta-coronavirus genus, HCoV-OC43 are among the known viruses that cause the common cold, but can also cause severe lower respiratory tract infections, including pneumonia in infants, the elderly, and immunocompromised individuals. Coronaviruses are enveloped viruses possessing a positive-stranded RNA genome with a length between 26.2 and 31.7 kb.

In order to overcome the low abundance of viral RNA and to take into account the constraints of routine sequencing we opted for an amplicon-sequencing approach with Illumina technology generating 300bp overlapping paired-end reads allowing to obtain a fragment of _˜500-550 bp after merging with casper [1]. A design of PCR primers pairs to cover the entire genome (around 30,600 bp) was performed using Primer-Blast onto a consensus sequence obtained from 47 HCoV-OC43 coronavirus full-sequenced genomes available in Genbank at that time. Primers were selected in conserved regions of the consensus sequence assuring enough overlapping between amplicon fragments for assembly of a single consensus genome after trimming of the primer sequences, leading finally to a set of 115 primers pairs.

Conventional NGS assemblers are not adapted to deal with a viral RNA population. Several tools dedicated to virus sequence assembling were developed, such as IVA (Iterative Virus Assembler)

*Speaker

[2]. However, they were not efficient in our case because of the difference of sequencing depth for each amplicon resulting from the amplicon-seq strategy which is not taking into account and leads to a broken up assembly (around 30 contigs were obtained with IVA for the HCoV-OC43 coronavirus genomes). In order to deal with the varying sequencing depth, we developed a complete protocol allowing to directly assemble virus genomes from amplicon sequencing data.

The first step consists in assigning each merged read to its amplicon of origin. For this assignment, we used the PCR primers as barcodes in a demultiplexing strategy using Cutadapt [3]. At the end of this step, the number of reads per amplicon is known: some amplicons were over-sequenced (up to 480000 reads) and some other were under-represented (from 0 to a few dozen of reads). For each amplicon, the presence of several sequence variants occurred due to the presence of a viral population. We postulated that the majority variant of the population corresponds to the most abundant read for each amplicon. A clustering strategy was used to identify the most abundant variant sequence for each amplicon using CD-hit [4] (clustering at 100% of identity). The most abundant read for each amplicon was then used as the representative fragment for the assembling step performed using cap3 [5].

Using this new protocol, we were able to obtain fully finished genomes for 7 out of the 13 clinical samples. For 3 other genomes we obtained 2 to 3 contigs covering more than 99% of the entire genome, with gaps that could be supplemented by Sanger sequencing. Some of the amplicon sequences obtained with our approach were checked by Sanger sequencing. Finally, the last 3 samples contained very low quantity of viral RNA explaining the poor sequencing results.

The analytical protocol introduced here is efficient, accurate and it allows virus sequencing in a clinical context for a routine use. As it is completely independent of a mapping step onto a reference sequence, it is able to detect highly variable sequences provided that the primer regions are conserved. This protocol was developed and used for the sequencing of HCoV-OC43 coronaviruses but can easily be used for other virus genomes, requiring only the design of the PCR primer set adapted to the studied virus.

References:

CASPER: context-aware scheme for paired-end reads from high-throughput amplicon sequencing. Sunyoung K. et al. BMC Bioinformatics 201415 (Suppl 9): S10 doi.org/10.1186/1471-2105-15-S9-S10

IVA: accurate de novo assembly of RNA virus genomes, Hunt M. et al. Bioinformatics, 2015 ;31(14):2374-6. doi: 10.1093/bioinformatics/btv120.

Cutadapt removes adapter sequences from high-throughput sequencing reads. Marcel Martin. EMBnet.journal dx.doi.org/10.14806/ej.17.1.200

CD-HIT: accelerated for clustering the next generation sequencing data. Limin Fu, et al. Bioinformatics, (2012), 28 (23): 3150-3152. doi: 10.1093/bioinformatics/bts56

CAP3: A DNA Sequence Assembly Program. Xiaoqiu Huang et al. Genome Res. 1999 Sep; 9(9): 868–877. PMID: 10508846

Keywords: sequencing, assembly, genome, virus, amplicon, seq, clinical, demultiplexing, clustering

Comparaison des approches d'étude du microbiote en contexte clinique : le séquençage 16S et le Whole Genome Sequencing

Aziza Caidi * , Denis Mestivier *

[1], Iradj Sobhani , Emma Bergsten

[1] Institut Mondor de Recherche Biomédicale – Responsable de plateforme de bioinformatique – 8 Général Sarrail, Créteil, France

Le microbiote intestinal joue un rôle important dans la santé de son hôte et des changements de sa composition (dysbiose) sont associés avec des états pathologiques (obésité, diabète, cancers, etc) [1] [2] .
Le nombre d'études analysant le microbiote a augmenté ces dernières années grâce au développement de technologies de séquençage à haut débit. Ces technologies ont révolutionné l'étude du microbiote puisque la majorité ($>$ 90%) des espèces microbiennes n'est pas cultivable par les techniques de culture de laboratoire actuelles [3]. L'amplification et le séquençage du gène de la sous unité ribosomale 16 (16S) reste la technologie la plus utilisée pour l'identification des microbiotes en métagénomique clinique [4]. Le séquençage du génome complet (WGS), alternative au 16S pour l'analyse du microbiote consiste à séquencer des régions chevauchantes du génome microbien en utilisant des amorces aléatoires permettant alors une meilleure identification de l'espèce [5]. Cependant cette approche coûte plus cher que la première et nécessite une grande couverture du génome pour une meilleure identification et compréhension du microbiote.

Dans le contexte clinique du cancer colo-rectal (CCR, un des trois cancers les plus fréquents), grâce à ces deux approches, différentes études sur cohorte ont mis en évidence des dysbioses associées à ce cancer.

Peu d'études ont comparé le microbiote obtenu par ces deux approches dans un cadre clinique et nous ne savons donc pas à quel point le séquençage 16S peut être une approximation de génome complet des bactéries et donc à quel niveau de fiabilité on peut se positionner. Or, une telle information est capitale car le 16S est l'approche de choix dans un cadre clinique au vu de son coût et de sa rapidité.
Une étude récente [6] conclut à une faible reproductibilité entre 16S et WGS sur une cohorte clinique de CCR. Nos résultats récents nuancent ces conclusions sur l'annotation taxonomique entre 16S et WGS avec des très bonnes corrélations pour un certain nombre de taxons. Ainsi, l'objectif de ce travail sera de : 1) comparer les signatures métagénomiques du CCR sur une cohorte de patients témoins et CCR séquencés **à la fois** en 16S et WGS et, 2) comparer ces

[*] Speaker

signatures à celles déjà publiées par Vogtmann et al [6] via une approche similaire. Nos résultats aideront à documenter les limites d'utilisation du 16S en contexte de métagénomique clinique.

Keywords: Microbiote, Métagénomique, 16S, WGS, Cancer colorectal, cadre Clinique

GDSCTools for mining pharmacogenomic interactions in cancer

Thomas Cokelaer *† 1

1 Hub Bioinformatique et Biostatistique - Bioinformatics and Biostatistics HUB – Institut Pasteur [Paris], Centre National de la Recherche Scientifique : USR3756 – C3BI, 25-28 rue du Docteur Roux, 75724 Paris cedex 15, France

Large pharmacogenomic screenings integrate heterogeneous cancer genomic datasets as well as anti-cancer drug responses on thousand human cancer cell lines. Mining this data to identify new therapies for cancer sub-populations would benefit from common data structures, modular computational biology tools and user-friendly interfaces. We have developed GDSCTools: a software aimed at the identification of clinically relevant genomic markers of drug response. The Genomics of Drug Sensitivity in Cancer (GDSC) database (www.cancerRxgene.org) integrates heterogeneous cancer genomic datasets as well as anti-cancer drug responses on a thousand cancer cell lines. Including statistical tools (analysis of variance) and predictive methods (Elastic Net), as well as common data structures, GDSCTools allows users to reproduce published results from GDSC and to implement new analytical methods. In addition, non-GDSC data resources can also be analysed since drug responses and genomic features can be encoded in standard formats (e.g. CSV files, dataframes, etc).

Keywords: drug discovery, oncology, pharmacology, machine learning

*Speaker

†Corresponding author: thomas.cokelaer@pasteur.fr

Multi-omics pan-cancer classification and biomarker design using machine learning

Aurélie Gabriel *† 1, Nicolas Alcala 1, James Mckay 1, Lynnette Fernandez-Cuesta 1, Matthieu Foll‡ 1

1 Genetic Cancer Susceptibility Group, Section of Genetics, International Agency for Research on Cancer – IARC(WHO) – France

Thanks to initiatives like The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC), scientists have access to different 'omics data generated from multiple cancer types. Several studies take advantage of these data to perform pan-cancer classification but often focus on one data category rather than combining multiple types of data (WXS, RNAseq, CNVs, methylation ...). A multi-omics approach could increase the performance of such classifiers, allow the identification of the most informative features for the classification of different cancer types and provide guidelines for biomarkers design.

In this study we used the TCGA database to access multiple types of 'omics data issued from different cancer types. The TCGAbiolinks R package [1] allowed us to download the somatic mutations data, the Copy Number Variations (CNV) data, the expression and methylation data. To retrieve fusion transcripts data we used the Tumor Fusion Gene Data Portal [2], a database providing the fusion transcripts detected in the TCGA samples. To reduce dimensionality we performed a pre-selection of the features based on prior biological knowledge of each cancer. For the features based on somatic mutations, only the significantly mutated genes in more than 5% of the samples were selected. For each sample and each gene a mutation score was computed, the higher the impact of the mutation on the protein is, the higher is the score value. Since it is known that the mutation burden as well as the distribution of each nucleotide change is different depending on the cancer type, the total number of mutations and the distribution of the six nucleotides changes in the selected genes were added to the features. The DoCM database (database of curated mutations) [3] provides a list of hotspots in cancers, we filtered the hotspots based on their frequency and kept as features the ones found in more than 5% of the samples. The fusion transcripts features were also filtered based on their frequency, the fusion transcripts found in more than 2% of the samples were selected. For the CNV features, based on GISTIC [4] results, we considered only the genes significantly mutated in more than 5% of the samples. The features based on methylation data are a list of the top hyper and hypo-methylated genes provided by O. Gevaert et al. [5] and the features based on expression data are a list of the most discriminative genes for pan-cancer classification, issued from the study of Y. Li et al. [6]. All data types combined, we gathered about 600 features.

Pan-cancer classification

[*] Speaker

[†] Corresponding author: gabriela@students.iarc.fr

[‡] Corresponding author: follm@iarc.fr

Based on these features, we first performed a pan-cancer multi-omics classification analysis based on 6515 patients from 17 cancer types using machine learning algorithms (support vector machines, random forest). The training and parameter tuning were done on a training and validation set using a 10 fold cross validation. An independent test set was saved to test and evaluate the performance using the F1-score, the harmonic mean of precision and recall. We compared the performance of classifiers based on the different molecular features: mutated driver genes, mutational burden, distribution of nucleotides changes, copy number variation, fusion transcripts, expression and methylation. We show that combining different types of omics data can improve the classification performance. Also, as previously shown, expression and methylation data lead to higher performances. However, this information is difficult to translate into the clinical setting. In this context, our data suggest that features derived from mutation, copy number and fusion transcript data can achieve a good performance in many cases, which could be more easily translated in a clinical context. Also several misclassification can be interpreted and biologically explained. Gynecologic cancers and breast cancers are for example difficult to distinguish from each other and lung squamous cell carcinoma (LUSC) are often misclassified in the lung adenocarcinoma (LUAD) cohort as well as in the head and neck squamous cell carcinoma (HNSC).

Classification of a given cancer type against any other

In a second application, using the same features, we aimed to distinguish one particular cancer type from any other and applied feature selection methods to identify the smallest number of features needed for this task. Identifying the most important features for the classification could be used to design biomarkers for diagnostic. This method was first applied to the small cell lung cancer (SCLC) data issued from George et .al [7] as proof-of-principle. In fact a biomarker based only on the two most frequently mutated genes, TP53 and RB1, has already been design and applied to SCLC early detection with a good sensitivity and specificity [8]. In our top features we find these two genes as well as other variables characteristic of SCLC: the number of mutations and the proportion of C> A mutations, associated with tobacco smoking. For this cancer type, an optimal performance can be reached using only the top two features: RB1 and the total number of mutations. The approach was also tested on the lung adenocarcinoma (LUAD) cohort from the TCGA. In contrast with the SCLC cases, LUAD cases have a greater number of significantly mutated genes which complicates the design of an optimal biomarker for diagnostic. This difficulty is reflected in our results, unlike the SCLC example more than 10 features are required to reach an optimal performance. As for the SCLC case, the list of top features provided contains biologically relevant features.

Conclusion

In a pan-cancer classification analysis, we confirmed that expression and methylation data lead to higher performances. However they are currently not easily applied in a clinical setting. We showed that using other 'omics data can in several cases lead to a good performance. The same data can also be used to identify the smallest number of features needed for the optimal classification of a given cancer type against any other. This type of classification allowed us to identify a small number of biologically relevant features that could be used to design biomarkers for diagnostic purpose.

References

A. Colaprico et al. 2016. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 44(8):e71

K. Yoshihara et al. 2015. The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene 34: 4845–4854

B.J. Ainscough et al. 2016. DoCM: a database of curated mutations in cancer. Nat. Methods 13, 806–807.

C. H Mermel et al. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 12:R41

O Gevaert et al. 2015. Pancancer analysis of DNA methylation-driven genes using MethylMix. Genome Biol 16:17

Y Li et al. 2017. A comprehensive genomic pan-cancer classification using the TCGA gene expression data. BMC Genomics 2017:18:508

J George et al. 2015. Comprehensive genomic profiles of small cell lung cancer. Nature 2017:524(7563):47-53

P Avogbe et al. Manuscript in preparation.

Keywords: Multi omics, Tumor classification, TCGA, Pancancer, Biomarkers design, Machine learning

REGOVAR, logiciel libre pour l'analyse de données de séquençage haut débit pour les maladies génétiques rares

Anne-Sophie Denommé-Pichon *† [1,2], Olivier Gueudelot [3,4], Jérémie Roquet [5], David Goudenège [3,4], Dominique Bonneau [3,4], Consortium Regovar‡

[1] Service de Génétique – Centre Hospitalier Régional Universitaire de Nancy – 54500 Vandœuvre-lès-Nancy, France
[2] Nutrition-Génétique et Exposition aux Risques Environnementaux (NGERE) – Institut National de la Santé et de la Recherche Médicale : UMRS954/UMRS1256, Université de Lorraine – Université de Lorraine, Faculté de Médecine, 9 avenue de la Forêt de Haye, 54500 Vandoeuvre-les-Nancy, France
[3] Laboratoire de génétique – CHU Angers – 49933 Angers, France
[4] Biologie Neurovasculaire et Mitochondriale Intégrée – Université d'Angers, Institut National de la Santé et de la Recherche Médicale : U1083, Centre National de la Recherche Scientifique : UMR6214 – UFR Sciences Médicales, 1 rue Haute de Reculée 49045 Angers Cedex01, France
[5] dnalyze.me – Regovar – 75013 Paris, France

1 Présentation

REGOVAR est un projet collaboratif, libre et ouvert de logiciel d'analyse de données de séquençage haut débit avec une interface graphique simple et conviviale pour les panels de gènes, l'exome et le génome (DPNI, recherche de SNV, CNV, SV...). Le projet est financé dans le cadre d'un appel d'offre du GIRCI des Hôpitaux Universitaires du Grand Ouest (HUGO, Angers, Brest, Nantes, Poitiers, Rennes et Tours) pour structurer les généticiens cliniciens, biologistes et bioinformaticiens impliqués dans le diagnostic moléculaire des maladies génétiques rares. Si la bioinformatique médicale appliquée au NGS permet aujourd'hui d'analyser avec succès un grand nombre de données au sein d'un CHU ou d'une région, elle souffre d'un manque de coordination à l'échelle nationale. REGOVAR vise à impliquer et fédérer les différentes communautés concernées, sans limites institutionnelles ou géographiques. Il se base exclusivement sur des technologies et des logiciels libres et gratuits, éliminant toute contrainte contractuelle et budgétaire.

2 Fonctionnalités

Après quelques années de développement, le déploiement de REGOVAR dans les centres intéressés a commencé pour une utilisation aussi bien en recherche qu'en diagnostic. REGOVAR permet le traitement de données génétiques, depuis la récupération des fichiers produits par les séquenceurs, quelle qu'en soit la technologie, jusqu'à la génération de rapports illustrés et de comptes-rendus d'analyse en passant par les contrôles de qualité, la détection, l'annotation, le filtrage, la priorisation et la visualisation de variants.

*Speaker

†Corresponding author: as.denomme@outlook.com

‡Corresponding author: contact@regovar.org

Son architecture client-serveur permet une utilisation depuis des ordinateurs de bureau sous Windows, Linux et macOS, via une interface graphique claire et intuitive conçue pour des généticiens n'ayant pas nécessairement de compétence spécifique en bioinformatique (filtres de variants enregistrables, simplification de la bioanalyse...). Le déploiement de REGOVAR est facilité par l'utilisation de packages et une configuration automatisée (via SaltStack).

Sa conception modulaire permet d'intégrer dynamiquement de nouveaux pipelines, qui peuvent être partagés au sein de la communauté, quelles que soient leurs dépendances, grâce à une encapsulation Docker. Ces échanges de pipelines permettront à terme l'harmonisation des bonnes pratiques avec des pipelines unifiés nationalement et validés par l'ANPGM.

REGOVAR intègre une base de données principale dimensionnée pour supporter aussi bien l'analyse de panels, d'exomes, que de génomes complets. Cette base est enrichie de données publiques telles que celles provenant de gnomAD et dbNSFP, ainsi que de données locales. Il permet le stockage d'informations concernant les patients (imagerie, consentements, liens familiaux, phénotype...) et offre la possibilité de filtrer les variants par phénotype par l'intermédiaire de Human Phenotype Ontology. L'analyse des variants est facilitée par un module de recherche connecté à des outils tels que PubMed et OMIM, ainsi qu'un accès direct aux outils en ligne comme Varsome.

Il est conçu pour s'adapter aux évolutions futures et laisse la possibilité de s'interfacer avec d'autres applications. À terme, des échanges de certaines informations anonymisées (fréquence des variants, pathogénicité des variants, phénotype) seront également possibles, selon une granularité laissée à la discrétion des généticiens.

3 Appel à collaboration
REGOVAR est également déployé dans le laboratoire de génétique du CHRU de Nancy et suscite l'intérêt d'autres acteurs comme le laboratoire de génétique du CHU de Montpellier. Le projet est ouvert à toute personne souhaitant apporter sa contribution : idées, intégration de pipeline, développement, test, documentation... Informations disponibles sur https://regovar.org/.

Keywords: NGS, exome, génome, opensource, libre, diagnostic, recherche

Mutational landscape of synchronous and metachronous breast cancer metastases through whole exome sequencing

Zakia Tariq * [1], Keltouma Driouch , Sylvain Baulande , Virginie Raynal , Virginie Bernard , François-Clément Bidard , Vanessa Benhamo , Ivan Bièche , Brigitte Sigal , Rosette Lidereau , Paul Cottu

[1] Institut Curie – Institut Curie – 26 rue dÚlm 75248 PARIS CEDEX 05, France

Background: Tumor progression is an evolutionary process associated with the accumulation of somatic genomic alterations. Even though metastasis is a main cause of cancer-related death and a challenging issue for the progress in targeted treatments, the molecular mechanisms of cancer metastasis remain poorly understood. Despite the advance in high throughput sequencing and the characterization of hundreds of breast tumor genomes, it is still unclear whether cancer genomes evolve by neutral processes or whether genetic alterations that favorable tumor progression are selected under pressures of cancer therapies or both.
To gain insights into the molecular processes driving breast cancer metastasis and to identify targetable genetic events associated with early and late stages of the disease, we studied the mutational profiles of a series of paired primary breast tumors and subsequent metastases collected as part of the clinical trial ESOPE meant to identify the "Changes in phenotype and genotype of breast cancers during the metastatic process and optimization of therapeutic targeting".

Methods: Matched primary breast tumors, first metastases and germline DNAs were obtained from 30 patients, divided in 2 groups; therapy-na´ive synchronous primary cancers and their distant metastatic lesions(n=9) and metachronous primary tumors and associated metastases that underwent first line adjuvant treatments (n=21). On Illumina platform, paired-end 100x100 whole exome sequencing was performed and three variant callers were used in parallel to predicted variations (Mutect1, HaplotypeCaller and UnifiedGenotyper). After Annovar annotation, variants with low 1000Genome frequency and absent in germline samples, were validated by IGV visualization and then considered as somatic variants. To characterize the drivers of breast cancer progression, we first annotated our variants using the referenced cancer genes database (Cancer Gene Census) and the litterature and we identified new metastasis drivers through Mutation Significance tool (MutSigCV). To determine copy number alterations, we used Facets and Sequenza tools and "Copynumber" R package to highlight recurrent variations. Samples purity and clonality was established by ABSOLUTE tool to evaluate metastasis divergence. Finally, an analysis of the signatures of mutational processes was performed with DeconstructSigs R package.

Results: In patients with synchronous disease, the mutational burden, the somatic variants

*Speaker

landscape, and the mutational signatures were strikingly similar in the primary tumors and their subsequent distant metastases. In contrast, patients with metachronous metastases exhibited a relative genomic divergence between the primary tumors and matched metastases, associated with the acquisition of metastases-specific alterations. The number of such alterations were independent of the time elapse between the diagnosis of primary tumors and the development of distant metastases. Analyses of mutated genes revealed that most of metastases-specific alterations identified in metachronous samples were already present in the primary tumors of patients with synchronous disease suggesting an early acquisition of the metastatic potential.

In conclusion, characterizing the genomic profiles of breast cancer metastases may have potential impact providing insight into the biology of the metastatic process as well as the mechanisms of treatment resistance. Ultimately, our findings might open new avenues for novel therapeutic strategies in metastatic breast cancer.

Keywords: Exome Sequencing, Breast Cancer Evolution, Metastases

VarAFT : Un système d'annotation et de filtration pour les données issues de séquençage haut débit.

Jean-Pierre Desvignes * [1], Marc Bartoli [1], Valérie Delague [1], Martin Krahn [2,1], Christophe Beroud [1,2], David Salgado [1]

[1] Universite d'Aix Marseille, Inserm, MMG INSERM-AMU U1251, Marseille, France (MMG) – Institut National de la Santé et de la Recherche Médicale - INSERM, Aix-Marseille Université - AMU – France
[2] APHM, Hopital Timone Enfants, Departement de Genetique Medicale, Marseille, France – Hôpital de la Timone [CHU - APHM] – France

Le séquençage haut débit (NGS) produit des centaines de millions de séquences par exome ou genome. L'analyse des données brutes jusqu'à l'identification des mutations n'est plus un frein mais l'étape d'annotation et de filtration reste encore délicate et cruciale dans l'identification des mutations candidates. Pour répondre à ce défi différents systèmes ont été développés ces dernières années mais aucun ne répond complètement aux besoins des utilisateurs tout en assurant la confidentialité des données.
Dans ce contexte nous avons développé VarAFT (Variant Annotation and Filter Tool – https://varaft.eu) qui est un logiciel multiplateforme écrit en JAVA permettant l'annotation, l'analyse et la filtration des mutations ainsi que l'analyse de couverture.

Le premier module permet l'annotation de fichiers VCF/gVCF grâce au système ANNOVAR implémenté et adapté aux différents systèmes d'exploitation. Ce module récupère également les prédictions d'UMD-Predictor (http://umd-predictor.eu) et de Human Splicing Finder (HSF - http://umd.be.hsf3/) via une API dédiée. L'annotation d'un exome de 60 000 variants ne prend que 8 minutes avec une configuration standard (4Gb RAM + 4 threads sans base de données optionnelles). Ce module permet également l'annotation de CNV (Copy Number Variation).

Le second module permet de combiner les fichiers annotés. Il peut être aussi bien utilisé dans une analyse familiale quel que soit le mode de transmission, que dans une analyse de cohorte ou que dans l'étude de cancers (Tumeur vs Normal). L'analyse peut être préalablement restreinte grâce à l'utilisation d'une liste de gènes ou d'un fichier BED. Ce module offre de nombreuses options de filtration permettant de réduire le nombre de mutations. Ces filtres utilisent les informations sur la nature des variants (RefSEQ et Ensembl) ; la fréquence d'observation (gnomAD,1000 génomes ...) ; les prédictions de différents outils (SIFT, PolyPhen, CADD, UMD-Predictor, HSF ...) ; des données liées au gène (OMIM, HPO, Gene Ontology, KEGG, Reactome, PID, GTeX). Il est également possible de créer sa propre base de données locale de mutations pour supprimer les variants récurrents. L'ensemble des variants résultant de l'analyse sont contenus dans un tableau dynamique à partir duquel l'utilisateur a simplement accès à l'ensemble des informations associées aux variants.

[*] Speaker

Différentes fonctionnalités ont été également ajoutées pour rendre l'utilisation de VarAFT plus convivial, comme la sauvegarde des filtres pour réutilisation en un clic, sauvegarde des données au format Excel ou bien dans le format VarAFT pour réutilisation. Il est également possible d'observer les variants d'intérêts sur le fichier BAM grâce à l'interconnexion entre VarAFT et IGV.

Pour faciliter la filtration de gros fichiers ou de nombreux fichiers, un système de traitement automatisé utilisant des filtres prédéfinis par l'utilisateur a été implémenté. Ce système permet de filtrer un génome de plus de 4 millions de variants en moins d'une minute.

Enfin le troisième module permet l'évaluation de la qualité de l'expérience. Il fournit la couverture de chaque transcrit, exon jusqu'au niveau de la base grâce à des tableaux et graphiques dynamiques qui peuvent être exportés et inclus dans des rapports. L'analyse de couverture est réalisée grâce à l'implémentation du système BEDtools également adaptés aux différents OS. Une analyse de couverture ne dure que 7 min pour tout un exome et moins de 30 secondes pour l'analyse d'un panel à partir d'un BAM d'exome.

VarAFT est aujourd'hui très utilisé tant pour la recherche que pour le diagnostic. Il a été téléchargé plus de 1200 fois (utilisateurs uniques source Google Analytics) dans 56 pays depuis Novembre 2016. Il a ainsi permis l'identification de nouveaux gènes et/ou de nouveaux variants impliqués dans différentes pathologies génétiques (Esteves et al, Turunen et al (2018), Jallades et al, Saultier et al, Elouej et al, Marquet et al, Cerino et al (2017), Galant et al, Lacoste et al, Miltgen et al (2016), Rapetti-Mauss et al 2015, Bartoli et al 2014).

VarAFT a également servi de modèle pour la création de la plateforme RD-Connect (Thompson et al 2014). Une publication est en cours de révision dans Nucleic Acid Research.
En conclusion VarAFT est un outil d'annotation et de filtration multiplateforme le plus complet accessible aux non bioinformaticiens et permet en peu d'étapes d'obtenir une liste réduite de mutations candidates facilitant ainsi la recherche et le diagnostic. Il peut être utilisé pour des panels de gènes, des exomes ou des génomes.

Keywords: Mutations, Génétique, NGS, Annotation, Filtration, Logiciel

Deducing cellular composition of complex airway epithelia from single cell RNA-seq data phenotyping

Marin Truchi * [1], Pascal Barbry† [1], Agnès Paquet‡ [1]

[1] IPMC – CNRS : UMR6097 – France

Complete authors list : Truchi M, Deprez M, Lebrigand K, Magnone V, Pons N, Arguel MJ, Cazareth J, Zaragosi LE, Leroy S, Marquette CH, Giovannini-Chami L, Paquet A*,Barbry P*

Introduction

Bulk RNA sequencing has been used in numerous studies. While highly informative, this approach does not discriminate between modifications caused by differences in cellular composition or by altered gene expression inside a homogeneous population of cells. More recent approaches, such as single cell RNA sequencing (scRNA-seq), can better document the cellular composition of a tissue and the specific expression profiles of each cell population. For cost and experimental matters, bulk RNA-seq still remains the most widely used technique in large comparative studies of patient cohorts or in routine clinical diagnostic. Combining the two approaches appears promising. Indeed, regression algorithms have been developed to infer cell proportions from bulk RNA-seq samples based on prior information derived from scRNA-seq data. These "deconvolution" methods can solve for each gene a system of equations by considering the expression in a bulk sample as a linear combination between its reference expression profiles in all present cell types and the relative proportions of cell types. Deconvolution represents an interesting *in silico* methods to mine cellular heterogeneity information.

The aim of our study is to apply this approach to epithelial cells of the respiratory airway, which play a central role in normal and pathological lung function. Firstly, we established the identity card of individual airway cells, isolated from nasal brushings or biopsies. Then, we assessed the performance of several deconvolution algorithms using simulated bulk RNAseq were the proportion of each cell type is known.

Methods

Deconvolution Tools

*Speaker

†Corresponding author: barbry@ipmc.cnrs.fr

‡Corresponding author: paquet@ipmc.cnrs.fr

We selected 2 deconvolution tools, CIBERSORT (Newman, *Nat Methods* 2015) and EPIC (Racle, *eLIFE* 2017), used in the literature to determine cell type proportions of bulk samples from reference gene expression profiles obtained by scRNA-seq analysis. In the CIBERSORT approach, which uses support vector regression, the sum of non negative coefficients values is forced to be equal to one. EPIC uses constrained least square regression and is designed to estimate the fraction of cells in bulk data that are uncharacterized in the reference.

Experimental data

Single cell gene expression profiles from nasal turbinates, nasal brushings and blood were processed on the 10x Chromium system using standard protocols, and sequenced on an Illumina NextSeq 500.

scRNA-seq analysis pipeline

All scRNAseq analysis steps were performed with the R package Seurat. Cells were first filtered according to quality control metrics such as the total number of UMIs per cell, the number of genes expressed per cell, or the fraction of mitochondrial genes. The data were then normalized by the total expression in each cell, multiplied by a scale factor, and log-transformed. Then, we applied unsupervised analysis to cluster the cells with similar gene expression profiles together. First, a principal component analysis was performed and the principal components containing most of the signal were used in the clustering step. Seurat clustering constructs a K-nearest neighbor graph based on the euclidean distance in PCA space, where cells with similar gene expression patterns are linked. A modularity optimization algorithm was applied to partition the cells. Results of the clustering were visualized using a t-SNE representation.

Cell type identification

A systematic ”one versus all” differential expression analysis was performed for each cluster. Genes with the most significant differences between their mean expression in a particular cluster and the rest of the cells were listed as specific markers. These markers were then used to assign a cell type to each cluster, according to their known biological relevance.

Processing of scRNA-seq reference matrix and bulk RNA seq samples

Deconvolution algorithms require building a reference matrix for each cell type. We selected very specific markers for each cell type (from 3-7 by cell types), with no overlap between cell type whenever possible. Then, we converted the scRNAseq UMI counts into CPMs, and computed the average CPMs for these genes in each of our single cell RNA clusters. These averages were then used as reference matrix in our deconvolution pipeline. Bulk RNAseq samples were converted into TPM before deconvolution.

Experimental validation strategy

Deconvolution algorithms performances were then tested on simulated data. Artificial bulk samples were generated by aggregating expression values of randomly sampled cells from an airway epithelium single cell dataset, using known proportions of each cell type. We also added expression values of uncharacterized melanoma cells to compare the algorithms performance to detect a fraction of "alien" cells.

Results

Single Cell RNAseq analysis

Our scRNAseq analysis allowed us to characterize 11 major cell populations in nasal biopsies: apical, basal, multiciliated cells, goblet cells, ionocyte-like cells, mast cells, macrophages, plasma cells, endothelial cells, NK cells and serous cells. These 11 populations were used as input in our deconvolution matrix.

Performance on simulated data

Both CIBERSORT and EPIC performed well in our simulation. Based on 20 independent simulations, we observed an average Spearman correlation of 0.95 for CIBERSORT, and 0.88 for EPIC. Both algorithms are highly sensitive to cell populations with specific and well expressed markers. For example, we were able to correctly assess a proportion of mastocytes down to 0.2% of the total cell population. A different situation was observed when cell markers overlapped with other cell types. An open question that remains regards the detection of unknown cell types, which was poor with both approaches. CIBERSORT algorithm does not support at all this feature. EPIC was able to correctly detect the presence of 10% of melanoma cells added to our simulated data, but was not sufficiently sensitive to detect the absence of some populations. This point is probably explained by the use of insufficiently specific gene expression markers.

Conclusion

We used single cell RNAseq data to build a catalog of gene expression profiles for 11 different cell types, including successive stages of airway epithelial cell differentiation, corresponding to the major cell types present in the airway epithelium. This catalog was then used to assess the performance of two deconvolution algorithms based on simulated data. Both algorithms performed well in our hands. The estimation of cell type proportion was better with Cibersort, but EPIC allowed the detection of unknown cell types, which can be very useful when studying gene expression data from uncharacterized samples. The estimated proportion of cells determined by our analysis pipeline can then be used as covariates to enrich differential expression analysis of clinical samples.

We will apply our deconvolution pipeline to bulk RNA-seq samples from a clinical study comparing gene expression in healthy and asthmatic airway epithelia. This will allow us to elucidate the influence of cell heterogeneity on gene expression in the different samples and gain new insights about the physiopathology of asthma.

Keywords: single cell RNA seq, airway epithelium, deconvolution

Skip-E : un système bioinformatique dédié au saut d'exon thérapeutique.

David Salgado* [1], Jean-pierre Desvignes [1], Marc Garibal † [1], Christophe Beroud‡ [1,2]

[1] Marseille medical genetics - Centre de génétique médicale de Marseille (MMG) – Aix Marseille Université : UMR$_S$1251, $InstitutNationaldelaSantéetdelaRechercheMédicale$: $UMR_S1251 - -FacultédeMédecine - Timone27, boulevardJeanMoulin13385Marseillecedex5, France$
[2] Assistance Publique - Hôpitaux de Marseille (APHM) – Institut National de la Santé et de la Recherche Médicale - INSERM – Direction générale AP-HM80, rue Brochier13 354 Marseille Cedex 5, France

Les mutations responsables des maladies génétiques chez l'homme peuvent avoir un impact sur les ARNm et/ou les protéines. Afin de restaurer une fonction protéique normale, diverses approches thérapeutiques ont été développées telles que **la thérapie génique** et la production de **protéines médicaments**. Ces approches complexes ont eu quelques succès mais restent limitées dans leurs applications. Parallèlement ont émergé des alternatives utilisant de petites molécules permettant d'interagir avec la machinerie cellulaire. De nombreux développements sont ainsi réalisés dans le cadre de **la translecture des codons** stop (interaction avec les ribosomes) et le **saut d'exon thérapeutique** (interaction avec le splicéosome). Cette dernière approche permet de restaurer le cadre de lecture rompu par différentes mutations (nonsense, insertions délétions hors phase, mutations d'épissage) et ainsi de produire une protéine tronquée quasi-fonctionnelle. Cette approche est aujourd'hui largement utilisée dans le cadre de la myopathie de Duchenne (DMD) ou différents essais cliniques encourageants sont en cours.
Elle repose sur l'utilisation **d'oligonucléotides antisens** (AON) modifiés chimiquement (*N3'-P5' phosphoramidate (NP), 2' fluoro-arabino nucleic acid (FANA), locked nucleic acid (LNA), phosphorodiamidate morpholino (PMO), cyclohexene nucleic acid (CeNA), Tricyclo-DNA (tcDNA), Peptide nucleic acid (PNA), ...*) qui se lient spécifiquement à l'ARN pré-messager du gène cible pour "tromper" la machinerie d'épissage en masquant un ou plusieurs exons du gène cible.

Cette stratégie rend théoriquement possible la modification de tous les transcrits, cependant le choix des AON est encore aujourd'hui empirique. Nous avons ainsi développé **Skip-E** (http://skip-e.geneticsandbioinformatics.eu/) afin de rationaliser le choix de ces molécules.

Skip-E est un système informatique constitué de deux composants logiciels :

- Une base de données PostGreSQL, **AonDBase**, qui permet la collecte des données publiées des AONs utilisés chez l'homme ou dans des modèles animaux ; de leurs cibles ; leurs

*Corresponding author: david.salgado@univ-amu.fr

†Speaker

‡Corresponding author: christophe.beroud@inserm.fr

contextes thérapeutiques et leurs efficacités. A ce jour, **AonDBase** contient 400 AONs ciblant 7 gènes (*SMN1*, *SMN2*, *DYSF*, *DMD*, *MYBPC3*, *ACVR1*, *P2RX3*) dans le cadre de de l'amyotrophie spinale (*SMN1* et 2), des dysferlinopathies (DYSF), des myopathies de Duchenne/Becker (DMD), des cardiomyopathies hypertrophiques et dilatées (*MYBPC3),* des fibrodysplasies ossifiantes progressives (*ACVR1*) et des cystites (*P2RX3*).

- Une interface utilisateur qui associe une interface web et un "*genome browser*" qui permet d'accéder aux informations de la base de données PostGreSQL **Skip-E**. Cette dernière contient l'ensemble des **AON de 15 à 50 bases** (∼5 milliards) pouvant s'hybrider aux différents exons, et régions introniques flanquantes, de l'ensemble des transcrits humains connus. Ces informations sont couplées aux **caractéristiques physico-chimiques** (séquence, masse moléculaire, température de fusion) et **structurales** (structure secondaire et stabilité de cette structure) et leurs **cibles** potentielles au niveau génomique (régions extrageniques, exons, introns). L'**efficacité** potentielle de chaque AON est évaluée en combinant sa spécificité (capacité à ne cibler que la région d'intérêt) et sa capacité à masquer des signaux d'épissage telles que les : points de branchement, sites d'épissage, signaux auxiliaires de type ESE (*Exonic Splicing Enhancer*).

En conclusion, **Skip-E** est un système bioinformatique dédié au saut d'exon thérapeutique. Il permet d'optimiser le choix de nouvelles molécules médicament et ainsi d'accélérer le développement de stratégies de saut d'exon thérapeutique chez l'homme, ouvrant la porte à la médecine personnalisée.
La disponibilité de nouvelles données expérimentales, chez l'homme et l'animal, permettra en outre, de définir un score prédictif d'efficacité des différents AON afin de simplifier la sélection de ces molécules.

Keywords: Skip, E, AonDBase, saut d'éxon, ADN, ARN, thérapie génique, protéines médicaments, restaurer le cadre de lecture, dysferlinopathies (DYSF), myopathie de Duchenne (DMD), oligonucléotides antisens (AON), web

Inter-cellular gene coexpression between microglia and oligodendrocytes during microglia activation in the developing brain

Nour Touibi * [1]

[1] Neuroprotection du Cerveau en Développement (PROTECT) – Assistance publique - Hôpitaux de Paris (AP-HP), Hôpital Robert Debré, Université Paris Diderot - Paris 7, Institut National de la Santé et de la Recherche Médicale : U1141 – Bâtiment Ecran/3ème étage Point Jaune - 48 Bd Serurier 75019 Paris, France

Background: Prematurity is a term for the broad category of neonates born at less than 37 weeks of gestation. Preterm birth is associated with neuroinflammation that leads to cerebral injury and hypomyelination. Encephalopathy of prematurity is characterized by oligodendrocyte maturation arrest, and reduced brain growth [1]. Microglia, the CNS macrophage plays a major role in neuroinflammation process through a poorly understood mechanism [2]. There is a paucity of knowledge regarding how the immature microglia and oligodendrocytes of the developing brain regulate their activation and maturation.
Methodology: To investigate the role of microglia in preterm brain development, we used a validated mouse model where encephalopathy of prematurity was induced by systemic interleukin-1β (IL-1B) administration [3]. Microglia and oligodendrocytes were isolated using Magnetic-activated cell sorting at several time points. Genome-wide gene expression data were generated from these isolated cells with 6 biological replicates for each condition. Microglia were sorted at 4 time points (i,e n=48 ; IL-1B/PBS ; P1/P5/P10/P45), and oligodendrocytes at two time points (i,e n=24 ; IL-1B/PBS ; P5/P10). These data were explored to identify transcriptional effect of the systemic IL-1B in microglia and oligodendrocytes but also the dynamics of gene programs during development. We performed differential gene expression analysis using the limma package [4] to compare systemic IL-1B exposure with control conditions and to examine transcriptional changes over time in both cell types. Functional analysis of the differentially expressed genes was performed using Gene set enrichment analysis, GSEA [5] applied genome-wide to the ranked list of gene scores (reflecting both the significance and the magnitude of expression changes). We tested for enrichment of differentially expressed genes in each set of genes of the MSigDB [6]. Gene set enrichment scores and significance level of the enrichment (NES, P-value, FDR) and enrichment plots were provided in the GSEA output format developed by Broad Institute of MIT and Harvard (permutations = 100,000).

Results: Microglial transcriptional inflammatory response is the most important at P5 (5 days following IL-1B administration) and characterized by upregulation of inflammatory response (GSEA for HALLMARK_INFLAMMATORY_RESPONSE gene set in microglia at P5: NES = 5.76, P-value < 10e-5) and by downregulation of cell cycle genes (GSEA for HALLMARK_E2F_TARGETS gene set in microglia at P5: NES = -7.72, P-value < 10e-5). Expression changes after systemic exposure to IL-1B in microglia are significantly correlated to

*Speaker

those observed in oligodendrocytes at P5 and P10 (Spearman correlation coefficient 0.95, P-value < 2.2e-16 at P5, and 0.91, P-value < 2.2e-16 at P10). Differential expression analyses of another mouse model where inflammation was induced by exposure of pregnant mice to polyriboinosinic-polyribobocytidilic acid (poly I:C) on E12,5 or E14,5, and microglia were isolated at P1 and P56 [7] showed no significant correlation with the microglial expression changes observed in the IL-1B model.

Conclusion and discussion: To conclude, our results suggest that the microglia transcriptional response to inflammation varies according to the mode and developmental stage of its induction, is not instantaneous and has remoted impact on the brain. The transcriptional changes induced by systemic IL-1B are highly correlated between microglia and oligodendrocytes, suggesting common regulation. Next steps of the project include gene co-expression-based analyses with topological measures to pinpoint gene master regulators of inflammation-associated modules in both cell types. A better understanding of their regulation in this model could allow therapeutic modulation to mitigate the brain injury of prematurity and other cerebral inflammatory conditions.

Bibliography:

G. Ball, J. P. Boardman, D. Rueckert, P. Aljabar, T. Arichi, N. Merchant, I. S. Gousias, A. D. Edwards, and S. J. Counsell, "The Effect of Preterm Birth on Thalamic and Cortical Development," *Cereb. Cortex*, vol. 22, no. 5, pp. 1016–1024, May 2012.

M. L. Krishnan, J. Van Steenwinckel, A.-L. Schang, J. Yan, J. Arnadottir, T. Le Charpentier, Z. Csaba, P. Dournaud, S. Cipriani, C. Auvynet, L. Titomanlio, J. Pansiot, G. Ball, J. P. Boardman, A. J. Walley, A. Saxena, G. Mirza, B. Fleiss, A. D. Edwards, E. Petretto, and P. Gressens, "Integrative genomics of microglia implicates DLG4 (PSD95) in the white matter development of preterm infants," *Nat. Commun.*, vol. 8, no. 1, p. 428, 2017.

G. Favrais, Y. Van De Looij, B. Fleiss, N. Ramanantsoa, P. Bonnin, G. Stoltenburg-Didinger, A. Lacaud, E. Saliba, O. Dammann, J. Gallego, S. Sizonenko, H. Hagberg, V. Lelièvre, and P. Gressens, "Systemic inflammation disrupts the developmental program of white matter," *Ann. Neurol.*, vol. 70, no. 4, pp. 550–565, 2011.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, no. 7, pp. e47–e47, 2015.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 43, pp. 15545–50, Oct. 2005.

A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdottir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, Jun. 2011.

O. Matcovitch-Natan, D. R. Winter, A. Giladi, S. Vargas Aguilar, A. Spinrad, S. Sarrazin, H. Ben-Yehuda, E. David, F. Zelada Gonzalez, P. Perrin, H. Keren-Shaul, M. Gury, D. Lara-Astaiso, C. A. Thaiss, M. Cohen, K. Bahar Halpern, K. Baruch, A. Deczkowska, E. Lorenzo-Vivas, S. Itzkovitz, E. Elinav, M. H. Sieweke, M. Schwartz, and I. Amit, "Microglia development follows a stepwise program to regulate brain homeostasis," *Science (80-.).*, vol. 353, no. 6301, p. aad8670-aad8670, Aug. 2016.

Keywords: neuroinflammation, microglia, oligodendrocytes, transciptomics, gene networks, prematurity

BioInfuse v3

Ewen Corre [1], Julien Fumey [1], Florence Jornod *† [1]

[1] Association des Jeunes Bioinformaticiens de France (RSG France - JeBiF) – ISCB SC – France

BioInfuse est un concours de vulgarisation lancé par l'association des Jeunes Bioinformaticiens de France (RSG France - JeBiF) lors de JOBIM 2016. Sous forme de courtes vidéos les participants ont pour objectif de rendre accessible la bioinformatique au grand public (adolescents et adultes non scientifiques).
Nous vous présenterons ici la vidéo gagnante de la 3ème édition de BioInfuse.

Keywords: jebif, concours vidéo, bioinfuse, vulgarisation

*Speaker

†Corresponding author: florence@jornod.com

Bioinformatics for diagnosis and clinical trials

Elodie Girard * [1,2,3], Choumouss Kamoun [1,2,3], Laetitia Chanas [1,2,3], Julien Romejon [1,2,3], Camille Benoist [4], Virginie Bernard [4], Julien Masliah-Planchon [4], Claire Morel [5], Gaelle Pierron [4], Celine Callens [4], Ivan Bieche [4], Maud Kamal [5], Christophe Le Tourneau [2,5,6], Philippe Hupe [1,2,3], Nicolas Servant [1,2,3]

[1] MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France – MINES ParisTech, PSL Research University – France

[2] INSERM, U900, F-75005 Paris, France – Inserm – France

[3] Institut Curie, PSL Research University, Mines Paris Tech, Bioinformatics and Computational Systems Biology of Cancer, INSERM U900, F-75005, Paris, France – Institut Curie, PSL Research University, Mines Paris Tech, Bioinformatics and Computational Systems Biology of Cancer, INSERM U900, F-75005, Paris, France – France

[4] Department of Biopathology, Institut Curie, PSL Research University, Paris, France – Institut Curie – France

[5] Department of Drug Development and Innovation, Institut Curie, Paris, France – Institut Curie – France

[6] Versailles-Saint-Quentin-en-Yvelines University, Montigny-le-Bretonneux, France – Université de Versailles-Saint Quentin en Yvelines – France

Next Generation Sequencing (NGS) has redefined the landscape of human molecular genetic testing by allowing the parallelization of a high number of patient analyses in a cost- and time-effective manner. The detection of genomic alterations is a key process of the precision medicine and bioinformatics has become an important component in the clinical laboratories, helping in the discovery and interpretation of the high volume of data generated by pan-cancer multigenes panels. Using the Low Input Dual Strand Truseq Custom Amplicon (TSCA) Illumina technology on a NextSeq500, tumors of 16 patients are sequenced every two weeks at the Institut Curie in order to identify cancer-related or druggable alterations in different contexts: i) for clinical diagnosis from formalin-fixed and paraffin-embedded (FFPE) tumors and ii) for the SHIVA02 clinical trial from frozen metastatic tumors. In this context, we developed a bioinformatics routine process that includes several levels of expertise (ranging from biologists and clinicians to computer scientists, data managers, functional analysts and developers). Our bioinformatics environment ensures the integration and the traceability of all data, as well as the processing and analyses of genomic data in real time (snvs, indels and copy number variations). After a rigorous quality control, a clinical report is delivered to the clinicians and biologists for the therapeutic decision.

Keywords: diagnosis, precision medicine, genomic alterations, cancer, clinical trial

[*] Speaker

Molecular characterization of rhabdoid tumors based on highthroughput and multi-omics data analysis

Mamy Andrianteranagna * [1]

[1] Institut Curie – Institut National de la Santé et de la Recherche Médicale - INSERM, Institut Curie, PSL Research University – 26 rue dÚlm 75248 PARIS CEDEX 05, France

Rhabdoid tumors are very rare and highly malignant childhood tumours that can arise from any part of the body. These tumors are characterized by the bi-allelic inactivation of the SMARCB1 gene which encodes for a key member of the SWI/SNF chromatin remodelling complex. Poor prognosis is the major issue of conventional therapy, which is essentially chemotherapy and/or radiotherapy based on the age of the patient and tumor developmental stage [1, 2]. By analysing multi-omics high-throughput data from human and mouse models, our main goal is to elucidate the molecular mechanisms underlying the development of rhabdoid tumors in order to identify new therapeutic strategies in the context of precision medicine.

Transcriptomics (gene expression microarray, RNAseq), epigenetics (DNA methylation array, DNA methylation bisulfite sequencing, ChIPseq) and genomics (whole exome sequencing) data were generated from human biopsy, tumor cell lines, and mouse models. These data were processed using pipelines developed by the Institut Curie Bioinformatics Platform. Gene expression profile, DNA methylation pattern, mutational landscape, and alternative splicing mechanisms are currently investigated. We also integrated public data for rhabdoid and other paediatric tumors in our analysis.
Based on DNA methylation and gene expression profiling, recent studies in rhabdoid tumors show that human intra-cranial rhabdoid tumors, named AT/RT (for Atypical/Teratoid Rhabdoid Tumors), can be clustered into 3 distinct molecular subgroups: TYR, SHH, and MYC [3, 4]. Gene expression from human AT/RT samples collected in our lab also show 3 molecular subgroups that correspond to published data. In addition, based on the nearest shrunken centroids approach [5], we identified a set of subgroup signature genes that will be used as a diagnosis tool in clinical applications to distinguish the 3 subgroups. Expression of selected genes will be validated on a set of independent rhabdoid tumor samples using the nanostring gene expression assessment technology.

Recently, our team generated intra-cranial rhabdoid tumors from mouse using conditional inactivation of SMARCB1 [6]. Intriguingly, samples from mouse rhabdoid tumor models are clustered into 2 subgroups based on gene expression patterns. The two subgroups are now referred to as neuronal and non-neuronal. In order to correlate murine subgroups with hu-

*Speaker

man subgroups, we applied different statistical approaches, including meta-analysis, batch effect correction with a linear model followed by hierarchical clustering and sample-to-sample correlation. Mouse models developed by other labs were also integrated in the analysis. All methods suggest that the neuronal subgroup correlates with the SHH subgroup, while the non-neuronal subgroup correlates with MYC. In addition, based on our current results, no murine rhabdoid tumor model was found to be correlated with the human TYR subgroup. Numerous studies suggested that members of the SWI/SNF complex are involved in pre-mRNA splicing mechanisms [7, 8]. In order to explore whether splicing mechanisms are deregulated in rhabdoid tumors, we compared the amount of canonical alternative splicing events observed in I2A cell line (a human rhabdoid tumor SMARCB1-deficient cell line) before and after SMARCB1 re-expression. Analysis was conducted using rMATS [9] and SGSeq tools using both annotation-based and de-novo approaches. According to our results, a general perturbation of alternative splicing in I2A SMARCB1 deficient cells could not be confirmed. The amount of each splicing event does not seem to be dependent on the presence/absence of SMARCB1. However, genes that showed differential splicing between the SMARCB1-deficient I2A and SAMRCB1-positive I2A could be identified and further study on tumor samples and other rhabdoid cell lines will be conducted to confirm if these genes may play a role in rhabdoid tumor development.

Keywords: rhabdoid tumors, highthroughput data, omics, pediatric tumor, SMARCB1, data integration

BaSaV: New Database for antibiotic resistance

Aurélien Birer * [1], Richard Bonnet * † [1,2,3,4]

[1] CNR "Résistance aux antibiotiques" (CNR) – Santé publique France – 3 boulevard Fleming, 25030 Besançon, France
[2] M2iSH, UMR 1071 Inserm/University of Clermont Auvergne, Clermont-Ferrand 63000, France – UMR 1071 – France
[3] INRA USC 2018, Clermont-Ferrand 63000, France – INRA USC 2018 – France
[4] Laboratory of Bacteriology, CHU Clermont-Ferrand, 63000, France – CHU Gabriel Montpied [Clermont-Ferrand] – France

Since the creation of the French National Reference Center (CNR) for antibiotic resistance, this CNR leads three mains missions: monitoring the spread of bacterial genes coding for antibiotics resistance in France, exploring underlying mechanisms and detecting any emerging phenomena in this domain. With the advent of Next Generation Sequencing (NGS) technologies, the method used now for identifying the antibiotic resistance mechanisms is the Whole Genome Sequencing (WGS) of resistant bacterial strains and the screening of genes and mutations involved in the resistance to antibiotics using a list of known determinants. Today, the number of bacterial strains sequenced by NGS sequencers is around 60 by week in the 4 CNR sites and the number of reference genes and mutations to antibiotic resistance in our database of knows determinants is more over 3000 entries in all antibiotic family.
To maintain the analysis service, organize the storage of data and to allow the development of news services, we have implemented a web application (DaSaV) and a database (BaSaV) to allow the upload of NGS data and meta data, analyze them and store the outputs in the same place to provide download filtered comprehensible results files and a rapid visualization on DaSaV.

The analysis of the data is based on home pipelines with one of which is an assembly/genotyping pipeline, an another is a rapid genotyping pipeline include ARIBA tool. The resistance database used in theses pipelines is our home resistance database initially build from the public databases : CARD, RestFinder and Arg-annot.

DaSaV host a novel home pipeline of bacterial strains typing to response of the rapid evolution of world typing and the increase of number genome in bacteria database to have a more sensitive classification of the CNR bacterial strains. This pipeline is based on core-Genome reference and phylogenetic Tree algorithm.
BaSaV via DaSaV is available for the CNR personnels in a first time and public in a later time.

*Speaker

†Corresponding author: rbonnet@chu-clermontferrand.fr

Keywords: Database, Resistance, Bacteria, CNR, Genotyping, Antibiotic, Web, Application

Omics data analyses and integration for precision medicine applied to advanced cancers

Arnaud Guille * [1], Quentin Da Costa [1], Séverine Garnier [1], José Adélaïde [1], Nadine Carbuccia [1], Anthony Gonçalves [1,2], Daniel Birnbaum [1], François Bertucci [1,2], Max Chaffanet [1]

[1] Centre de Recherche en Cancérologie de Marseille (CRCM) – Aix Marseille Université : UM105, Institut Paoli-Calmettes : UMR7258, Institut National de la Santé et de la Recherche Médicale : U1068, Centre National de la Recherche Scientifique : UMR7258 – 27 bd Leï Roure, BP 30005913273 Marseille Cedex 09, France

[2] Institut Paoli Calmettes (IPC) – Fédération nationale des Centres de lutte contre le Cancer (FNCLCC) – 232, boulevard Ste Marguerite 13009 Marseille, France

Because cancer is a heterogeneous and complex disease, tumor molecular profiling has become a fundamental component of precision oncology, enabling the identification for each patient of genomic alterations in genes and pathways that could be targeted therapeutically. To define clinically actionable mutations, array CGH (aCGH) and next-generation sequencing (NGS) assays are main approaches to establish individual molecular portraits of tumors that oncologists will use to predict the most appropriate treatment for a given patient.

The amount and heterogeneity of generated data associated with molecular predictive medicine represent new challenges for the bioinformatics field. Indeed, this requires efficient systems to integrate data generated by different sources and to ensure the security of such sensitive data. Moreover, traceability of analyzed samples and reproducibility of analyses are required in a clinical setting.

Here, we present a bioinformatic workflow from the raw data (NGS, aCGH) to molecular reports used in the context of the precision medicine clinical trials at the Paoli Calmettes Institute (PERMED 01 study - ClinicalTrials.gov, NCT02342158; PANDORA BC-BIO study - ClinicalTrials.gov, NCT01521676) and applied for more than 500 advanced cancers. The sequencing pipeline was developed using Snakemake and Conda. Snakemake was used to build reproducible and flexible analysis pipeline and ensure traceability of the samples. Snakemake coupled with Conda was used to build an isolated environment including the hundred of softwares needed for the whole workflow. To characterize clinically-actionable gene mutations and putative druggable pathways, this pipeline was used to identify gene copy number aberrations (CNAs) and somatic mutation status by using respectively whole genome aCGH and targeted NGS of 559 genes or whole exome sequencing (WES). From the sequencing data, we also extracted additional information such as mutational signatures, mutational load and micro-satellites instability, which can be useful for therapeutic choices and diagnosis.

All these information were then stored in a mysql database driven by a django web application.

[*] Speaker

This driven web application database guarantees the security of the data and provides an elegant interface for researchers and oncologists who desire to retrieve corresponding alterations. Finally, a pdf molecular report was generated as a support for weekly molecular tumor board meetings to help oncologists in their treatment strategies.

Keywords: NGS, Sequencing, aCGH, Copy number, Precision medicine, Snakemale, Conda, Django, Analysis, Integration, Database

MobiCNV : a simple and rapid method for detecting Copy Number Variants in Illumina gene panel, clinical exome and exome data

Henri Pégeot [1], Charles Van Goethem [2], Gema Garcia-Garcia [1], Reda Zenagui [1], Delphine Lacourt [1], Olivier Ardouin [1], Mireille Claustres [1], Michel Koenig [1], Mireille Cossée [1], Anne-Françoise Roux [1], David Baux * [1]

[1] Laboratoire de génétique des maladies rares, EA7402, Université de Montpellier, Laboratoire de génétique moléculaire, CHU de Montpellier, Montpellier France (LGMR, CHU Montpellier) – CHU Montpellier, Université de Montpellier – France

[2] Laboratoire de biologie des tumeurs solides, CHU de Montpellier, Université de Montpellier, Montpellier France – CHU Montpellier, Université de Montpellier – France

The identification of small DNA events (substitutions, small insertions/deletions) has been tremendously improved during the last 3 years but it is still challenging to effectively and easily identify larger events such as deletions/duplications of more than one exon. Sophisticated algorithms have been developed based on the depth of coverage of the regions of interest (ROIs), however, if useful, they are often complex to operate and rely on large pre-computed data sets. We have been using a method for several years in our laboratory based on normalized depth of coverage comparison between samples and regions of a single Illumina run. We present here MobiCNV (https://github.com/mobidic/MobiCNV), a simple and rapid python script implementing this algorithm that can be run with only basic bioinformatics knowledge on UNIX or Microsoft Windows platforms. The script computes as input csv or tsv files summarizing the depth of coverage for each ROI obtained either directly from Illumina regular pipelines Local Run Manager® or MiSeq Reporter® or from a slightly modified samtools bedcov command. The more samples the run contains (at least 4, ideally between 8 and 20), the more sensitive the method is. MobiCNV performs more efectively with runs having an average sequencing depth above 100X and is therefore well suited for gene panel analyses in clinical context or for exomes with sufficient coverage. This algorithm also relies on the uniformity of the sequencing experiment and consequently requires good quality data. The method is particularly effective for multi-exon events. The output is a single Excel spreadsheet which contains several worksheets. The most useful one is the summary sheet, which presents all the regions containing a predicted event surrounded by the neighbouring regions for comparison. Moreover, MobiCNV includes a gender prediction module, based on the analysis of reads mapping on the X chromosome which ultimately produces a table summarizing the gender predictions of each sample. It is worth noting that sexual chromosomes and autosomes are treated separately by the algorithm. MobiCNV can also focus on a list of genes of interest provided as a simple text file (e.g. for large panels including multiple pathologies). MobiCNV is a simple script which efficiently analyses CNVs from gene panel and exome data, and that can easily be implemented as part of any analysis

*Speaker

pipeline.

Keywords: NGS, CNV, Copy Number Variants, Illumina sequencing

MPA, a free, accessible and efficient pipeline for SNV annotation and prioritization for NGS routine molecular diagnosis.

Kevin Yauy [1], David Baux *† [1], Henri Pégeot [1], Charles Van Goethem [2], Charly Mathieu [1], Thomas Guignard [3], Raul Juntas Morales [1], Delphine Lacourt [1], Martin Krahn [4,5], Vilma Lehtokari [6], Gisèle Bonne [7], Sylvie Tuffery-Giraud [8], Michel Koenig [1], Mireille Cossée [1]

[1] Laboratoire de génétique des maladies rares, EA7402, Université de Montpellier, Laboratoire de génétique moléculaire, CHU de Montpellier, Montpellier France (LGMR, CHU Montpellier) – CHU Montpellier, Université de Montpellier – France

[2] Laboratoire de biologie des tumeurs solides, CHU de Montpellier, Université de Montpellier, Montpellier France – CHU Montpellier, Université de Montpellier – France

[3] Plateforme Recherche de Microremaniements Chromosomiques, Hôpital Arnaud de Villeneuve, CHU de Montpellier, Faculté de Médecine Montpellier-Nîmes, Université de Montpellier, Montpellier, France – CHU Montpellier, Université de Montpellier – France

[4] Universite d'Aix Marseille, Inserm, GMGF INSERM-AMU UMRS910, Marseille, France – Institut National de la Santé et de la Recherche Médicale - INSERM – France

[5] APHM, Hopital Timone Enfants, Departement de Genetique Medicale, Marseille, France – Hôpital de la Timone [CHU - APHM] – France

[6] The Folkhalsan Institute of Genetics and the Department of Medical Genetics, Haartman Institute, University of Helsinki, Helsinki, Finland – Finland

[7] Sorbonne Universites, UPMC Univ Paris 06; INSERM U974; Center of Research in Myology; Institut de Myologie, Paris F-75013, France – Institut National de la Santé et de la Recherche Médicale - INSERM, Université Paris-Sorbonne, Université Paris-Sorbonne – France

[8] Laboratoire de Genetique des Maladies Rares, EA7402, Universite de Montpellier, Montpellier, France – Université de Montpellier – France

One of the main limitations of massively parallel sequencing for molecular genetic diagnosis is the interpretation of the huge amount of generated data. In the case of myopathies and muscular dystrophies, another major issue is to efficiently predict the pathogenicity of the many variants identified in large genes, especially the little known *TTN* gene. We propose the MoBiDiC Prioritization Algorithm (MPA) that gives a variant prioritization score based on the curated interpretation of previously reported variants, biological assumptions, and splice and missense predictors to prioritize all single nucleotide variant (SNV) types. We validated our approach by comparing the sensitivity and specificity of MPA and of prediction tools in dbNSFP using a dataset composed of *DYSF*, *DMD*, *LMNA*, *NEB* and *TTN* gene variants extracted from expert-reviewed and ExAc databases. MPA obtained the best annotation rate for missense and splice variants. As MPA aggregates results from several predictors, errors by individual prediction tools are counterweighted, thus improving the sensitivity and specificity of missense and splicing variant prediction. We propose a sequential use of MPA, beginning with the selection of the

*Speaker

†Corresponding author: david.baux@inserm.fr

variants with the highest MPA scores, followed, in the absence of candidate variants for the pathology, by inclusion of variants with lower scores. We provide scripts and documentation for free academic use and a validated annotation pipeline to prioritize SNVs from a VCF file (https://github.com/mobidic/MPA).

Keywords: NGS, SNV annotation, molecular diagnosis, muscular dystrophies

Développement d'un pipeline bioinformatique en oncogénétique clinique: sélection des outils, validation de seuils et exigence des normes

Quentin Da Costa * [1]

[1] Centre de Recherche en Cancérologie de Marseille (CRCM) – Aix Marseille Université : UM105, Institut Paoli-Calmettes : UMR7258, Institut National de la Santé et de la Recherche Médicale : U1068, Centre National de la Recherche Scientifique : UMR7258 – 27 bd Leï Roure, BP 30005913273 Marseille Cedex 09, France

Pour faire face à l'accroissement des volumes (+ 15% par an), de la complexité (diversification des panels de gènes) et de la réduction des délais d'analyse, le NGS a supplanté dans de nombreuses situations la méthode de séquençage Sanger qui cependant reste le gold standard. L'accréditation de cette approche, en particulier d'un point de vue bioinformatique pose de nouveaux défis. Nous présentons l'expérience du Laboratoire d'Oncogénétique Moléculaire (LOGM) de l'Institut Paoli-Calmettes (IPC), dans le cadre des analyses constitutionnelles et tumorales. La mise en place du NGS entraîne d'importants besoins en méthodes bioinformatiques. Ce poster expose la mise en place d'une **structure d'analyse bioinformatique à double pipeline** au LOGM (IPC), démarrant à la sortie du séquenceur (fichiers *.fastq) et aboutissant aux tableaux des variants annotés et à leur stockage.

La norme encadrant les activités des laboratoires de biologie médicale ISO-15189 a pour but d'assurer la fiabilité des processus mis en place, la traçabilité, et l'amélioration constante des processus.

Dans le cadre de la détection de variants en Oncogénétique, la fiabilité des résultats est constituée par :

1) La certitude de l'absence de variant en cas de résultat négatif ;

2) La certitude de l'existence du variant en cas de résultat positif.

1) Pour répondre au premier cas, une double détection des variants, très appréciée des auditeurs Cofrac, a été mise en place. En effet, nous faisons fonctionner en parallèle deux programmes : Sophia DDM (Sophia Genetics) et CLC Biomedical Genomics Workbench (Qiagen).

Les variants détectés par les deux pipelines sont comparés, intégrés et formatés dans un rapport Excel par des scripts (Python) permettant le développement dans un environnement Windows et leur utilisation par l'ensemble du personnel du laboratoire, avec les sécurisations requises..

*Speaker

2) **Le seuil minimal de détection** pour l'analyse des tissus fixés est difficile à définir. Il n'existe ni référence absolue, ni argument irréfutable. Pourtant, il s'agit d'un élément essentiel pour le dossier de validation de méthode.

En tenant compte de cela, dans le deuxième cas, dans le cadre des analyses d'ADN tumoral issu de tissu fixé, nous avons déterminé notre seuil minimal de fréquence des variants en modélisant la distribution des fréquences de 12 000 variants de notre laboratoire. Cela nous permet de nous abstenir d'une sélection arbitraire du seuil.
De plus, nous développons une base de données qualité (SeQual, sequencing quality), permettant l'analyse automatisée des données de séquençage et leur stockage dans une base de données. Une interface accessible sur navigateur web permettra de tracer les variables de contrôle qualité sur un intervalle de temps donné, permettant l'identification de dysfonctionnements sur le long terme. L'ensemble des variants détectés et annotés par le LOGM est stocké dans cette même base, ce qui en fait un outil exhaustif pour tout laboratoire d'Oncogénétique.

Keywords: Oncogenetics, Diagnostic, Clinic, Oncology, Cancer, Norme, Mutation, Gene, Variant

Application pour une lecture et une gestion facilitées des données brutes générées lors de l'analyse d'un panel CFTR par NGS

Laetitia Gaston * [1], Julie Tinat [1], Raphaelle Campait [1], Benoit Arveiler [2], Patricia Fergelot [2], Marie-Pierre Reboul [1]

[1] CHU Bordeaux, Service de Génétique Médicale – CHU de Bordeaux Pellegrin [Bordeaux] – France
[2] CHU Bordeaux, Service de Génétique Médicale – CHU de Bordeaux Pellegrin [Bordeaux], INSERM U1211, université de Bordeaux – France

La lecture des données brutes générées lors de l'analyse d'un panel par séquençage de nouvelle génération nécessite l'utilisation successive de plusieurs outils bioinformatiques appropriés. Ce présent travail est né d'un réel besoin de faciliter la lecture du NGS dans le cadre du diagnostic moléculaire de la mucoviscidose. Il résulte de la nécessité au sein du laboratoire de rationaliser la gestion et l'archivage de nos données et d'améliorer la traçabilité de nos analyses. Aussi nous avons développé une application qui intègre dans une seule interface graphique l'ensemble des informations relatives aux données générées par le séquenceur mais aussi des informations provenant de différents logiciels d'analyse appliqués aux données.

Dans notre laboratoire, l'analyse du gène *CFTR* repose sur un panel AmpliSeq$^{\text{TM}}$ de Life Technologies couplée au séquenceur PGM (Ion Torrent/Life Technologies). L'analyse des données brutes se fait grâce à l'utilisation successive de trois logiciels : Torrent Suite v.5.6, Ion Reporter$^{\text{TM}}$ v.5.2 et SeqPilot v.4.3.0.

L'application a été développée en python avec la bibliothèque graphique pyQt4. Il s'agissait d'y intégrer toutes les informations nécessaires provenant de ces 3 logiciels à partir de leurs fichiers de sortie : le .tsv d'Ion Reporter, le .txt de SeqPilot et le .xls de la Torrent suite, ainsi que le fichier bam du patient.

L'application peut prendre en compte dans son analyse une base de données SQL locale et évolutive de mutations et de polymorphismes caractérisés. Cette base de données a été implémentée à partir des variants du gène *CFTR* caractérisés par la communauté scientifique mais aussi à partir des variants identifiés au sein du laboratoire.

L'analyse réalisée par l'application permet d'extraire différents types de données. Tout d'abord le numéro de l'ADN du patient et les données d'identification du run au cours duquel a été analysé l'ADN et de l'automate sur lequel a été réalisée l'analyse. Puis l'ensemble des variants communs aux logiciels SeqPilot et Torrent suite sont extraits. Les variants spécifiques à chacun des trois logiciels sont aussi présentés séparément. Pour terminer la notification de l'identification comme hotspot par le logiciel Ion Reporter est reportée pour chaque variant.

*Speaker

Pour chaque variant un certain nombre de champs ont été ajouté afin de faciliter leur interprétation et le suivi de la validation dudit variant : parmi ces champs, le premier permet une interprétation préliminaire par comparaison avec la base de données associée à l'analyse, le deuxième permet le suivi de la demande de validation par séquençage Sanger et le troisième le résultat de cette analyse. Enfin, les deux derniers champs additionnels permettent au biologiste et au technicien de rentrer un commentaire. Les informations de couvertures minimale et maximale par base sont calculées pour chaque amplicon afin de vérifier sa couverture. Afin de permettre la traçabilité de l'‘analyse effectuée, il est possible de renseigner la date de validation et le nom des techniciens et biologistes ayant assurés la vérification technique ainsi que la validation biologique. Enfin, un commentaire général peut également être rattaché au rapport d'analyse.

A partir de cette interface il est possible d'éditer un rapport détaillant la liste des séquences Sanger à faire pour confirmer les variants présents identifiés par le biologiste. Enfin, chaque analyse peut être sauvegardée au format pickle afin de pouvoir la rouvrir ultérieurement dans l'application et suivre l'évolution de l'analyse en cours. Une fonction d'export, au format pdf, de l'ensemble des données nous permet l'intégration dans notre logiciel de gestion de laboratoire.

Dans l'objectif de maintenir à jour une base de données, les variants d'intérêt peuvent être sélectionnés et conservés pour les futures analyses afin d'aider à l'interprétation.
En conclusion, l'utilisation de cette application " tout-en-un " permet de gagner du temps dans l'analyse. Elle permet en éditant une liste de travail d'optimiser l'organisation des vérifications par le séquençage Sanger. Elle constitue un outil de traçabilité de l'analyse en permettant d'associer des biologistes et techniciens ainsi qu'en notifiant les dates de vérification et validation. Mais aussi elle facilite l'importation des résultats dans notre logiciel de gestion de laboratoire. Elle offre la possibilité de créer une base de données de variants évolutive pour l'aide à l'interprétation des analyses suivantes. Finalement elle nous aide au quotidien dans notre activité diagnostique au sein du laboratoire.

Keywords: génétique, application, variants, comparaison, interprétation, séquençage ciblé moyen débit, diagnostique, mucoviscidose

A centralized web interface for genomic medicine labs

Gaelle Vilchez *† [1], Pierre-Antoine Rollat-Farnier [1,2], Sylvain Picard [1], Nicolas Chatron [2,3], Audrey Labalme [2], Mathilde Di-Filippo [4,5], Damien Sanlaville [2,3], Pierre Blanc [6], Thomas Simonet [1,7,8], Claire Bardel‡ [1,9,10]

[1] Cellule bioinformatique de la plateforme NGS du CHU de Lyon – Hospices Civils de Lyon : Cellulebioinformatique de la plateforme NGS du CHU de Lyon – Groupement Hospitalier Est, 59 bd Pinel, 69677 BRON CEDEX, France

[2] Service de Génétique des Hospices Civils de Lyon – Hospices Civils de Lyon : Servicede Génétique – Groupement Hospitalier Est, 59 bd Pinel, 69677 BRON CEDEX, France

[3] Centre de recherche en neurosciences de Lyon, équipe TIGER – CNRS : UMR5292, Université Claude Bernard Lyon 1, Université de Lyon, INSERM U1028 – Institut des Épilepsies IDEE, 59 Boulevard Pinel, 69500 BRON, France

[4] Service de Biochimie et Biologie Moléculaire Grand Est, UF Dyslipidémies – Hospices Civils de Lyon : Servicede Biochimie et Biologie Moléculaire Grand Est, UF Dyslipidémies – Groupement Hospitalier Est, 59 bd Pinel, 69677 BRON CEDEX, France

[5] Cardiovasculaire, Métabolisme, Diabétologie et Nutrition (CarMeN) – Institut National de la Recherche Agronomique - INRA, Université Claude Bernard - Lyon I, Institut National des Sciences Appliquées (INSA) - Lyon, Université de Lyon, Inserm U1060 – Faculté de Médecine Lyon-Sud 165 Chemin du Grand Revoyet BP12 - 69921 OULLINS Cedex -France, France

[6] Plateforme de séquençage NGS du CHU de Lyon – Hospices Civils de Lyon : Plateformede séquençage NGS - CHU Lyon – Groupement Hospitalier Est, 59 bd Pinel, 69677 BRON CEDEX, France

[7] Institut NeuroMyoGène (INMG) – CNRS : UMR5310, Université Claude Bernard - Lyon I, Université de Lyon, INSERM U1217 – 8 Avenue Rockefeller, 69008 Lyon, France

[8] Centre de biotechnologie cellulaire (CBC) – Hospices Civils de Lyon : Centrede biotechnologie cellulaire – Groupement Hospitalier Est, 59 bd Pinel, 69677 BRON CEDEX, France

[9] Laboratoire de Biométrie et Biologie Evolutive - équipe biostatistique - santé (LBBE) – Université Claude Bernard Lyon 1, Université de Lyon, CNRS : UMR5558 – 162 avenue Lacassagne, 69003 Lyon, France

[10] Service de Biostatistique - Bioinformatique, CHU Lyon – Hospices Civils de Lyon : Servicede Biostatistique - Bioinformatique – 162 avenue Lacassagne, 69003 Lyon, France

In the last few years, clinical Laboratories were required to perform an increasing number of high throughput sequencing tests for diagnosis. The Lyon University Hospital (LUH) started its NGS routine diagnostic activity 4.5 years ago with semiconductor sequencers (PGM and Proton). In the following years, 14 teams of the LHU grouped on a single platform have progressively acquired 4 Illumina sequencers (three NextSeq 500 and one MiSeq) and generated more than 750 runs with this technology.

The increasing number of tests, both for somatic and constitutional diagnosis, makes it necessary to automate the bioinformatic analysis meanwhile controlling quality and traceability of the results. and to implement tools that facilitate variant interpretation for geneticists. The LUH

*Speaker

†Corresponding author: gaelle.vilchez@chu-lyon.fr

‡Corresponding author: claire.bardel@univ-lyon1.de

Bioinformatics Group has thus developed a user interface coupled to a database for complete control of the NGS data analysis process, from the raw data files of the sequencer (.bcl) to variant classification and generation of a clinical report. During the various analysis steps, all the information is stored in the database, which enables users to: 1) Organize the sequencing runs of the different teams into projects, 2) View and validate the sequencing runs 3) Launch bioinformatic analyses using an in-house pipeline manager named Papillyon, 4) Filter, sort and classify variants in order to identify variations responsible for the patient's phenotype, 6) Manage diagnostic confirmation techniques with automation of PCR primer design and finally 7) Generate a final report for clinical care.

Built in close collaboration with geneticists, the application was optimized to ensure fluidity of the variant analysis process. Public and commercial databases used for the annotation (eg GnomAD (http://gnomad.broadinstitute.org/), OMIM (https://www.omim.org/) and HGMD Pro (http://www.hgmd.cf.ac.uk/ac/index.php)) are implemented within the application. Semi-automatic variant classification is performed according to the the ACMG / ANPGM guidelines (American College of Medical Genetics and Genomics / National Association of Molecular Genetics Practitioners).Finally, every step of the process is traced following ISO15189 requirements. This application offers automation and flexibility to the geneticists' requests. A computer script retrieves all the information related to a given project present in the database and thus makes it possible to perform a bioinformatic analysis adapted to 1) the type of mutation sought (single nucleotide variation, structural variation including copy number variants), 2) sample origin (constitutional, somatic or mitochondrial DNA) 3) library preparation protocol . Within a year, the application allowed to perform the analysis of more than 150 sequencing runs for more than 4000 patients. Further developments are ongoing, such as the support of transcriptomic data or data analysis automation of non-invasive prenatal screening for trisomies 13, 18 and 21.

Keywords: NGS, Database, Interface, Variants, Analysis, Hospital, Diagnostic, Automation, Annotation, Illumina, Traceability

SHAMAN : a shiny application for quantitative metagenomic analysis

Amine Ghozlane * [1], Stevenn Volant† [1]

[1] Bioinformatics and Biostatistics Hub, C3BI, USR 3756 CNRS, Institut Pasteur, Paris, France (C3BI) – Institut Pasteur de Paris – France

Background:
Quantitative metagenomics is broadly employed to identify genera or species associated with several diseases. These data are obtained by mapping the reads of each sample against operational taxonomic units (OTU) or a gene catalog. SHAMAN (shaman.pasteur.fr) [Quereda et al. 2016 (PNAS)] was one the first web application that allowed to clinician and biologist to perform an interactive analysis of quantitative metagenomics data with a dynamic-interface dedicated to the diagnostic and to the differential analysis. The interface integrates the experimental design (association of sample to one or several conditions), the statistical process for differential analysis and a real-time visualisation system.

Methods:

SHAMAN is based R, Shiny and DESeq2. The analytical process is divided into four steps : count matrix/annotation submission, normalisation, modelisation and visualisation. The count matrix is normalised at the OTU/gene level using the DESeq2 normalisation method and then, based on the experimental design, a generalised linear model is applied to detect differences in abundance at the considered taxonomic level.

Results:

Two years later, we can see a great interest from the metagenomics community with 3 publications using SHAMAN, 74 active users per month (1430 unique visitors since first publication) and 514 downloads of the docker application. Several developments are now available since first publication : a full automatized bioinformatic workflow for target metagenomics, a docker application for local installation on windows/mac/linux (aghozlane/shaman on docker hub) and several new visualizations for the quality check and a phylogenetical perspective of the abundance.

Keywords: metagenomics, differential analysis

*Speaker

†Corresponding author: stevenn.volant@pasteur.fr

FI-MICS test: innovative tool for Cardiac enriched lncRNAs measurement

Sami Ait Abbi Nazi *† [1], Eric Schordan‡ [1], Sabrina Danilin [1], Hueseyin Firat [1]

[1] Firalis SA – Firalis – 35, rue du Fort 68330 Huningue, France

In recent years, long non-coding RNA (lncRNAs) have emerged has a new type of non-coding RNA and many studies have shown their potential as powerful biomarkers in various pathologies such as cancer.

In contrast to other non-coding RNA such as miRNA or snoRNA, lncRNAs lack strong whole sequence conservation across different species but rather appear to contain short, highly conserved elements. Despite only a few lncRNAs having been shown to be biologically relevant and functionally annotated, there's growing evidence that the majority of them are likely to be functional. While the exact function of most IncRNAs remains unknown, they have been implicated in various biological processes, mainly relating to transcriptional, post-transcriptional and epigenetic regulation. The majority of IncRNAs to date, that are functionally characterized, are believed to regulate developmental processes. However, recent profiling of the mice cardiac transcriptome, after myocardial infarction in mice cardiac tissue, has shown their role in controlling mature tissue as well as the relevance of their expression level in cardiac pathologies.

Cardiac disorders such as coronary artery disease (CAD), acute myocardial infarcation (AMI) and heart failure (HF) are leading causes of mortality and morbidity in the world and cardiac toxicities such as those induced by drugs and drug candidates are the most important cause of drug withdrawal. Thus, there is a very important unmet medical need for diverse types of biomarkers for assessing cardiac function, including but not limited to diagnosis, prognosis, monitoring of drug effects and diseases activity. Several cardiac pathologies remain still incurable or need less aggressive and more personalized treatment. lncRNAs represent a novel family of targets useful for these diagnostic and therapeutic applications in the cardiovascular area. The present invention relates to lncRNAs that are cardiac enriched and described for the first time in human cardiac tissues. They represent good therapeutic and diagnostic candidates for cardiac related disorders. Like miRNA, lncRNAs can be released from the original tissue into the body circulation. Thus, such markers may be detected in body fluids like whole blood or plasma, which facilitates their use in clinics.

In collaboration with Lxembourg Institue of Health (LIH), Centre Hospitalier Universitaire Vaudois (CHUV) and Cardinal Stefan Wyszynski Institute of Cardiology, a deep-sequencing of cardiac biopsies with several conditions (control patients, ischemic cardiomyopathy and dilated cardiomyopathy) was performed to identify lncRNAs, which are enriched in cardiac tissue. Three

*Speaker

†Corresponding author: sami.nazi@firalis.com

‡Corresponding author: eric.schordan@firalis.com

ways are used for the identification: (a) known transcripts, (b) orthologs from the mouse model and (c) discovery of novel transcripts. 3228 lncRNAs were identified as cardiac tissue enriched lncRNAs. These lncRNAs may be involved in different pathophysiological events pertaining to cardiac function and represent a potential target for therapeutic approaches.

In the aim to detect these 3238 lncRNA, we present a targeted sequencing preparation kit: the FI-MICS test. This FI-MICS test is designed to measure lncRNAs in peripheral blood (Paxgene, Serum, Plasma) and biopsies. A comparison between a total RNA-seq protocol and the use of FIMICS test was performed. Using this FI-MICS test approach, a higher coverage is obtained with 20x less reads, enabling large multiplexing of samples.
With this FI-MICS kit, Firalis aim to develop an innovative in vitro diagnostic (IVD) test to assess the risk of developing heart failure (HF) after acute myocardial infarction (AMI), including lncRNAs in a predictive model. This model will be further validated within the ongoing HEARTLINC clinical study based on 300 AMI patients launched in June 2017.

Keywords: heart failure, NGS, lncRNA

RABIOPRED, an innovative theranostic tool to assist clinicians select an optimal anti-TNF alpha biological therapy for Rheumatoid Arthritis Patients

Eric Schordan* [1], Sami Ait Abbi Nazi †‡ [1], Sabrina Danilin [1], Matthieu Coq [1], Madah Mehdi [1], Hueseyin Firat [1]

[1] Firalis SA – Firalis – 35, rue du Fort 68330 Huningue, France

Background: TNF alpha blockers form 2nd line treatment choice for Rheumatoid Arthritis (RA) patients. Up to 30% of RA patients do not respond to TNF alpha blockers for unknown reasons, causing a significant impact on patients' outcome and healthcare industry. Therefore, there is an unmet need for a tool to predict treatment response that could help clinicians to choose an optimal treatment for RA patients.
Objectives: By using Immuno-Detect, an innovative targeted gene sequencing panel of 2155 mRNA targets associated with immune-inflammatory pathways, we aimed to develop an algorithm, RABIOPRED, that predicts non-response to TNF alpha blockers.

Methods: Paxgene samples obtained at baseline from 68 patients na´ive to TNF alpha blockers were directly profiled without extraction with Immuno-Detect panel on HTG EdgeSeq platform, a combination of a nuclease protection assay & next generation sequencing (NGS). Patients were treated with Infliximab, Etanercept or Adalimumab and disease activity score was measured based on DAS28 score at 3 months. Response to treatment was assessed by categorizing the patients according to EULAR response criteria. Gene combinations were selected using variable importance score (VIS). Predictive modeling performance was evaluated using the area under the curve (AUC) and confusion matrix.

Results: Analytical validation of Immuno-Detect panel shows a very high reproducibility on Paxgene and extracted RNA samples with correlation factor of 0.975 and 0.96 respectively. In paxgene samples, among 2155 genes, 1172 mRNAs are significantly expressed with a mean CV of 9.77% (976 mRNAs and mean CV of 11.98% for RNA). Most expressed target represent only 5% of the total reads and only 20 targets are reaching 1% of total reads showing a very well balanced panel. HTG data were validated against qPCR and shows a correlation of 0.69 and 0.8 for Paxgene and extracted RNA respectively. Performance of our predictive model shows an AUC of 0.905 with 0.88 accuracy. Our algorithm predicts non-responders to TNF alpha blockers with the sensitivity of 0.78 and positive predictive value of 0.91. This algorithm will be further validated within the ongoing RABIOPRED Proof-of-Performance study (ClinicalTrials.gov Identifier: NCT03016260) based on 720 patients treated by anti-TNF alpha drugs (5 originators

*Corresponding author: eric.schordan@firalis.com

†Speaker

‡Corresponding author: sami.nazi@firalis.com

& 3 biosimilars) launched in December 2016.
Conclusions: We are showing that Immuno-Detect panel accurately measures mRNA expression using HTG-EdgeSeq NGS platform. This panel can be further used to build signatures to predict TNF alpha blocker's non-response. Preliminary performance of the current assay shows that it can efficiently predict treatment response to anti-TNF alpha biologicals. The algorithm will be later on validated in a multi-centric proof-of-performance clinical study.

Keywords: rheumatoid arthritis, RABIOPRED, NGS, HTG EdgeSeq

Rencontre autour de l'Enseignement en BioInformatique en France (REBIF 2018)

Morgane Thomas-Chollier * [1], Stéphane Le Crom [2], Jacques Van Helden [3]

[1] Institute of Biology at the Ecole Normale Superieure (IBENS) – INSERM 1024 CNRS 8197 – 46 rue d'Ulm 75005 Paris, France
[2] Sorbonne Université, Univ Antilles, Univ Nice Sophia Antipolis, CNRS, Evolution Paris Seine - Institut de Biologie Paris Seine (EPS - IBPS) – Université Pierre et Marie Curie - Paris 6 – France
[3] Aix-Marseille University (AMU) – Aix-Marseille University – Aix-Marseille Université, Parc Scientifique et Technologique de Luminy, Case 928, 13288 Marseille cedex 9, France +33 4 91 82 87 12, France

Deux ans après la première édition de "Rencontre autour de l'Enseignement en BioInformatique en France", la Société Française de Bioinformatique (SFBI) a réuni à nouveau les acteurs de la formation diplômante en Bioinformatique au niveau national, afin de leur permettre d'échanger et de créer un véritable réseau de formations.
Ces deuxièmes Journées REBIF on eu lieu les 31 mai et 1 juin 2018 à Massy (http://www.sfbi.fr/rebif2016), en accueillant une trentaine de personnes (plus de 20 formations seront représentées). REBIF est soutenu par la Société Française de Bioinformatique (SFBI) et l'Institut Français de Bioinformatique (IFB).

Une des actions de REBIF est d'établir la liste exhaustive des formations diplômantes en Bioinformatique de tous niveaux (DUT, Licences, Master, Ingénieur, DU), sous forme de fiches formations standardisées, avec des mots-clefs permettant de caractériser chaque formation et ainsi faciliter l'orientation des étudiants. Ces fiches, présentant le paysage actuel des formations et de leur spécificités, ainsi qu'une carte de France des formations, seront rendues publiques sur le site de la SFBI, et viendront compléter le listing maintenu par l'association des Jeunes Bioinformaticiens de France (JeBiF) (https://jebif.fr/fr/bioinformatique/les-formations/).

Cette édition a été couplée à la "Reunion autour des Métiers de la Bionformatique" (MetBIF), qui a eu lieu le jour précédent. Une restitution du paysage des métiers, et des compétences professionnelles a été faite, afin de les mettre en regard des compétences à acquérir. Deux ateliers ont été proposé, l'un autours d'un socle commun de compétences en Bioinformatique, et l'autre autour de l'enseignement des bonnes pratiques professionnelles. Enfin, l'unité d'enseignement Meet-U a été présentée, afin de mettre en valeur les actions d'enseignements inter-formations.
Alors que pour les futurs étudiants, le paysage des formations reste assez flou, et qu'il n'est pas évident pour eux de s'orienter vers la formation la plus adaptée à leurs profils, le poster présenté incluant la carte des formations et leurs mots-clefs, ainsi que les grandes conclusions des ateliers permettra d'offrir une vision d'ensemble des formations françaises en bioinformatiques.

[*] Speaker

Keywords: enseignement, formation

JeBiF fête ses 10 ans !

Florence Jornod [1], Julien Fumey [1], Alexandre Borrel [1], Magali Michaut [1]*†

[1] JeBiF – ISCB SC RSG-France – France

Créée en 2008, l'association des Jeunes Bioinformaticiens de France (JeBiF) fête cette année ses 10 ans. Grâce à plus de 100 volontaires, JeBiF a pu pérenniser son action au fil des ans et s'établir comme un élément important du paysage bioinformatique français, tout en faisant partie d'un réseau international de Regional Student Groups (RSGs) ayant en commun leur lien avec le Student Council de l'International Society for Computational Biology (ISCB).
Depuis l'édition de Nantes en 2009, JeBiF a organisé chaque année un évènement en marge de JOBIM, créant du lien au sein de la jeune génération et diffusant de l'information sur les formations, métiers et carrières dans la bioinformatique. Développer son réseau et collecter de l'information a aussi été possible grâces notamment aux "P'tits Dejs JeBiF", aux "JeBiFs Pubs" et aux "TOBIs".

L'action de JeBiF s'étend au-delà des jeunes bioinformaticiens de France : i) des évènements de vulgarisation ont été organisés, notamment à l'occasion de la fête de la science, pour familiariser un public plus large à la bioinformatique, et plus récemment avec le concours de video de vulgarisation Bioinfuse ; ii) des évènements internationaux ont également été organisés en collaboration avec d'autres RSGs (Pays-Bas, Luxembourg, Belgique).

JeBiF a pu compter sur des collaborations fructueuses avec la SFBI, Bioinfo-fr.net et la CJC, ainsi que sur le soutien régulier du GDR BIM, des chefs de groupes qui permettent à leurs étudiants de participer aux initiatives de JeBiF, et de vous tous.

Un grand merci et vivement les 10 prochaines années !

Keywords: association, communauté, JeBiF, ISCB, jeunes, formation, métier, information

*Speaker

†Corresponding author: magali.michaut.rsg@gmail.com

Association des Jeunes Bioinformaticiens de France

Florence Jornod *† [1], Benvegnu Benvegnu-Sallou [1], Stéphanie Chevalier [1], Athénaïs Vaginay [1], Aurélien Béliard [1], Julien Fumey [1], Victor Grentzinger [1], Sylvain Léonard [1], Thibaut Payen [1], Hugo Pereira [1]

[1] Association des Jeunes Bioinformaticiens de France (RSG France - JeBiF) – ISCB SC – France

L'association des Jeunes Bioinformaticiens de France (RSG France - JeBiF), fondée en 2008, a pour mission de structurer la jeune communauté bioinformatique au niveau local, national et international. Elle constitue la partie française du réseau étudiant de la Société Internationale de Bioinformatique (International Society for Computational Biology Student Council – Regional Student Group France)
JeBiF organise divers évènements tels que des Tables Rondes dans des masters, des JeBiF Pubs et Tables Ouvertes en Bioinformatique, son Workshop annuel en marge de JOBIM, elle participe aux Bioinformations, et encourage également la découverte de la bioinformatique par le grand public à travers divers projets de vulgarisation tels que le concours Bionfuse, des activités periscolaires ou encore sa participation régulière à la fête de la science.

Cette année est particulière pour l'association car elle fête ses 10 ans. Nous vous présenterons ainsi les événements principaux ayant marqué la vie de l'association ces dernières années.

Keywords: communauté, association, iscb

*Speaker

†Corresponding author: florence@jornod.com

Taking genomics into the clinic: whole genome sequence of 13,000 patients with a rare disorder.

Karyn Megy *† [1,2], Rutendo Mapeta [1,2], Salih Tuna [1,2], Olga Shamardina [1,2], Sri Deevi [1,2], Hannah Stark [1,2], Christopher Penkett [1,2], Kathleen Stirrups [1,2], Lucy Raymond [1,3,4], Willem Ouwehand [1,2,3,5]

[1] NIHR BioResource (NIHR BR) – Cambridge University Hospitals, Cambridge Biomedical Campus, Cambridge, United Kingdom

[2] Department of Haematology, University of Cambridge – Cambridge Biomedical Campus, Cambridge, United Kingdom

[3] The Wellcome Trust Sanger Institute [Cambridge] – Hinxton, Cambridge CB10 1SA, UK, United Kingdom

[4] Department of Medical Genetics, Cambridge Institute for Medical Research – University of Cambridge, Cambridge, United Kingdom

[5] NHS Blood and Transplant – Cambridge Biomedical Campus, Cambridge, United Kingdom

Introduction

There are at least 7,000 rare diseases and 1 in 17 people are affected by a rare disease at some point in their life. The journey for patients and their close relatives for reaching a diagnosis lasts on average 2.2 years. Until recently, analysis has been limited to the sequencing of a few candidate genes, and obtaining a conclusive diagnosis was often not achieved. In 2013, the NIHR BioResource commenced one of the Rare Diseases pilots for the 100,000 Genomes Project and now 13,027 DNA samples from patients with rare diseases and their close relatives have been analysed by whole genome sequencing (WGS). Generating the WGS results and the associated metadata preceded the interpretation by multidisciplinary teams (MDTs) in the context of phenotype data, to generate clinical-standard reports for the referring clinicians.

Materials and Methods

Recruitment, Sequencing and Processing

A selection of patients, drawn from fifteen groups of rare diseases, were approved for enrolment. They and their relatives were consented for participation at 56 UK hospitals, and at hospitals in Europe and the United States. Samples of blood or DNA were collected, accompanying clinical and laboratory phenotypes were encoded using the Human Phenotype Ontology (HPO), and entered in the NIHR BioResource database [1]. Quality controlled DNA was analysed by WGS at Illumina Cambridge Limited, UK and single nucleotide variants (SNVs), short insertions and deletions (InDels) and structural variants (SVs) were called using, respectively, Isaac [2], Manta [3], and Canvas[4]. The variant results were transferred to the University of Cambridge's High

*Speaker

†Corresponding author: km369@medschl.cam.ac.uk

Performance Computing facility and further quality control was used to discard low quality variant calls. Gender was inferred based on the number of reads mapped to chromosomes X and Y relative to those that mapped to autosomes, and ethnicity and relatedness were estimated using Principal Component Analysis (PCA).

Variant prioritisation and filtering strategy

In order to identify putative causal variants, we prioritised variants based on (i) their minor allele frequency (MAF) in control populations (gnomAD [5]) with MAF < 1:1000 for putative novel causal variants and MAF

References

Westbury *et al.* Genome Med. **2015** Apr 9;7(1):36. PMID 25949529.

Raczy *et al.* Bioinformatics. **2013** Aug 15;29(16):2041-3. PMID 23736529.

Chen *et al.* Bioinformatics. **2016** Apr 15;32(8):1220-2. PMID 26647377.

Roller *et al.* Bioinformatics. **2016** Aug 1;32(15):2375-7. PMID 27153601.

Lek *et al.* Nature. **2016** Aug 18;536(7616):285-91. PMID 27535533.

McLaren *et al.* Genome Biol. **2016** Jun 6;17(1):122. PMID 27268795.

Stenson *et al.* Hum Genet. **2017** Jun;136(6):665-677. PMID 28349240.

Richards *et al.* Genet Med. **2015** May;17(5):405-24. PMID 25741868.

Landrum *et al.* Nucleic Acids Res. **2016** Jan 4;44(D1):D862-8. PMID 26582918.

Firth *et al.* Am J Hum Genet. **2009** Apr;84(4):524-33. PMID 19344873.

Turro *et al.* Sci Transl Med. **2016** Mar 2;8(328):328ra30. PMID 26936507.

Stritt *et al.* Blood. **2016** Jun 9;127(23):2903-14. PMID 26912466.

Arno G *et al.* Am J Hum Genet. **2017** Feb 2;100(2):334-342. PMID 28132693.

Gr´af *et al.* Nat Commun. **2018** Apr 12;9(1):1416. PMID 29650961.

Keywords: clinical, genomics, rare, disorders

Systems biology
Functional genomics

A map of direct TF – DNA interactions in the human genome

Marius Gheorghe [1], Geir Kjetil Sandve [2], Aziz Khan [1], Jeanne Chèneby [3,4], Benoit Ballester [3,4], Anthony Mathelier *† [1,5]

[1] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway – Norway
[2] Department of Computer Science, University of Oslo, 0316 Oslo, Norway – Norway
[3] INSERM, UMR1090 TAGC, Marseille, France – Centre de Recherche Inserm – France
[4] Aix-Marseille Université, UMR1090 TAGC, Marseille, France – Aix-Marseille Université - AMU – France
[5] Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway – Norway

A map of direct TF – DNA interactions in the human genome

Marius Gheorghe1, Geir Kjetil Sandve2, Aziz Khan1, Jeanne Chèneby3,4, Benoît Ballester3,4, Anthony Mathelier1,5

1 Centre for Molecular Medicine Norway (NCMM), University of Oslo, Oslo, Norway

2 Department of Informatics, University of Oslo, Oslo, Norway

3 INSERM, UMR1090 TAGC, Marseille, France

4 Aix-Marseille Université, UMR1090 TAGC, Marseille, France

5 Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway

Background

Transcriptional regulation is a biological mechanism essential to cell growth and differentiation, and is achieved through interactions between proteins and the DNA. This process is primarily controlled by transcription factors (TFs), which are key proteins specifically binding to the DNA at *cis*-regulatory regions to control the rate of transcription of RNAs. Specifically, TFs recognize their binding sites (TFBSs) through a complex interplay between nucleotide/base readout and DNA shape readout1. Hence, the identification of TFBSs genome-wide is a critical step towards understanding how gene expression regulation works.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) represents the most popular experimental assay to identify the genomic regions, so called ChIP-seq peaks, where TFs bind

*Speaker

†Corresponding author: anthony.mathelier@ncmm.uio.no

to DNA *in vivo*2. Unfortunately, it has recurrently been shown that ChIP-seq experiments are prone to generate ChIP-seq artifacts, and delineating bona fide bound regions from experimental noise is still an ongoing problem3–6. The ever-increasing number of publicly available ChIP-seq data sets provides an unprecedented opportunity to develop and evaluate computational tools designed to infer the precise locations of the TFBSs within ChIP-seq peaks by combining both computational and experimental evidences of direct TF – DNA interactions. While TFBSs are traditionally modeled through position weight matrices (PWMs), which represent the probability of each nucleotide to be present at each position within TFBSs, more advanced computational methods have been recently developed to incorporate nucleotide dependencies, variable spacing, and DNA conformation in the models. Assessment of these recent TF binding models on their capacity to predict TFBSs within ChIP-seq peaks highlighted that a one-fits-all model for TFBS prediction is not applicable7,8. In this work, we planned to automatically detect direct TF – DNA interactions from ChIP-seq data, increasing the confidence in predicted TFBSs.

Results

ChIP-eat, a framework to automatically detect direct TF – DNA interactions in ChIP-seq peaks

We have developed ChIP-eat, a uniform ChIP-seq data processing pipeline, from raw ChIP-seq reads to high confidence direct TF – DNA interactions. In the last update of the ReMap database9, we have uniformly processed, on the hg38 version of the human genome, more than 3,000 ChIP-seq datasets available from public sources, covering a total of 496 distinct ChIP'ed proteins. From this set, TF binding profiles were available in the JASPAR database10 for 232 TFs, corresponding to 1,983 ChIP-seq datasets. As TFBSs are expected to be enriched at the 'peak-summit' positions of ChIP-seq peaks11, where the highest number of ChIP-seq reads map, we assessed four classes of models for TFBS prediction by computing enrichment of predicted TFBSs at those positions. The four different prediction models evaluated were PWMs (optimized with DiMO12), the transcription factor flexible models (TFFM7), binding energy models, and DNA-shape-based models8, varying from simple to complex. Subsequently, to define the set of predictions that are most likely to represent the direct binding of the TFs to the DNA, we used a non-parametric, data driven entropy-based algorithm13 to automatically define thresholds on the motif scores (representing how similar a DNA sequence is to the modeled TFBSs), and on the distance of the predicted TFBSs from the peak-summit (representing how proximal it is with respect to the genomic location where the most ChIP-seq reads aligned). We considered that bona fide TFBSs were more likely to be located close to the peak-summits with high motif scores. The entropy-based algorithm allows to compute the two thresholds that define an enrichment zone where predicted TFBSs present strong both experimental and computational evidences of direct TF – DNA interactions. When applied with DiMO-optimized PWMs, ChIP-eat predicts direct TF – DNA interactions (TFBSs) that cover $> 4\%$ of the human genome.

A posteriori validation of the TFBSs predicted by ChIP-eat

To validate our set of TFBS predictions a posteriori, we first assessed the binding affinity of TF – DNA interactions derived experimentally via protein binding microarray (PBM)14. We found significantly higher PBM median intensity (p-value < 0.05, Mann-Whitney U test) for the set of DNA sequences within our enrichment zones than for the sequences outside for 81% of the 249 datasets accounting for 57 TFs for which PBM data was available.

Next, we observed that the p-values of the ChIP-seq peaks (from the MACS2 peak caller) containing a TFBS predicted within the enrichment zones were significantly lower than the rest of the peaks, which are not predicted to contain a direct TF – DNA interaction, for 96% of the 1,983 datasets. To further assert the confidence of the predicted TFBSs, we applied ChIP-eat to ChIP-seq peaks from two other peak callers: HOMER and BCP. We observed that ChIP-seq peaks containing a direct TF – DNA interaction predicted by ChIP-eat were more likely to be called by all three peak callers than the non-reproducible peaks between peak callers.

Finally, we applied our methodology to two ChIP-exo datasets (for ESR1 and FOXA1) and found that > 90% of the peaks that we predicted to be derived from direct TF – DNA interactions were also predicted as such by the ChExMix tool15, which has been specifically developed to characterize protein-DNA binding subtypes in ChIP-exo data. Using Jaccard similarity index, we obtained Jaccard similarities of 64 and 69 for ESR1 and FOXA1, respectively. This suggests that ChIP-eat, primarily developed for ChIP-seq data, which is more noisy and less precise than ChIP-exo, is able to extract binding regions derived from direct TF – DNA interactions when employed on ChIP-exo data.

Altogether, these results confirmed the predictions of TFBSs defined from the enrichment zones as very likely to be bona fide direct TF – DNA interactions.

High-occupancy target regions are likely to be derived from ChIP-seq artifacts or indirect TF binding

High-occupancy target (HOT) regions are genomic regions where ChIP-seq peaks for a large number of TFs are observed16. These regions are widespread throughout the genome, but they were shown to be depleted for the canonical motifs recognized by the ChIP'ed TFs, and their functional significance is yet unclear5. We used our set of high quality TFBS predictions to confirm that HOT regions were depleted of direct TF – DNA interactions. Indeed, we found that ChIP-seq peaks that are likely ***not*** derived from direct TF – DNA interactions (i.e. without a TFBS in the enrichment zones) were located significantly more often in HOT regions than peaks derived from direct TF – DNA interactions (p-value < 2.2e-16, hypergeometric test). This further suggests that despite the fact that HOT regions present an increased number of ChIP-seq peaks, they are not derived from direct binding of the TF to the DNA, which is in line with previous studies showing that HOT regions might be ChIP-seq artifacts5.

Direct TF – DNA interactions predicted by ChIP-eat reveal co-binding TFs

We used our ChIP-eat predictions to reveal TFs with co-localized TFBSs in the human genome. Pairs of genome-wide co-localized TFs were inferred from all the combinations of the 232 available TFs (i.e. 26,796 pairs tested). This was achieved by comparing the geometric mean of the genomic distance between TFBSs for each pair of TFs to the mean genomic distance expected by chance (given the complete set of predicted TFBSs). In total, 150 pairs of TFs accounting for 112 distinct TFs were associated with TFBSs that were significantly (FDR < 0.05) co-occurring genome-wide. To confirm these predictions of co-localized TFs on the human genome, we extracted protein-protein interaction networks from GeneMANIA17 and found that 82% of our predicted TF pairs were already known to physically interact. The remaining pairs could serve as entry points for future studies. Taken together, these results confirm the efficacy of such an extensive collection of high confidence TFBS predictions.

UniBind, a map of direct TF – DNA interactions in the human genome

We collected our complete set of TFBS predictions from each prediction model, as well as the optimized models themselves and the original ChIP-seq peaks, and made them publicly available through the UniBind database, encompassing the entire collection of ChIP-seq datasets uniformly processed through the ChIP-eat pipeline. UniBind provides an interactive web interface with easy browsing, searching, and downloading for all our predictions. For instance, users can search for predictions for specific TFs, cell lines, and conditions. Finally, we used the tool CREAM18 to predict *cis*-regulatory modules (CRMs) in the human genomes. Specifically, CREAM was applied to our set of predicted TFBSs for each cell line, and on the complete set, to predict genomic regions that correspond to clusters of direct TF – DNA binding events. The CRMs are publicly available in UniBind for further studies.

Conclusion

We have uniformly processed about two thousand publicly available ChIP-seq datasets from raw ChIP-seq reads to high quality TF – DNA binding interaction events on the human genome. To predict these TFBSs, we used a non-parametric, entropy-based algorithm to automatically find thresholds on TF binding model scores and distance to peak-summit that defined an enrichment zone. The accuracy of our predictions was a posteriori validated using *in vitro* assays, ChIP-exo, and multiple peak calling algorithms. The results show that our predictions are supported by strong experimental and computational evidences for direct TF – DNA interactions. This set of TFBSs allowed us to confirm that HOT regions are likely to be derived from ChIP-seq artifacts or indirect binding of the TFs to the DNA. We predicted TFs that are found to be co-localized on the human genome and that could be cooperate to achieve specific functions. Finally, the complete set of predictions is made publicly available through the UniBind database and web-interface, in an effort to provide the community with a remarkable collection of high quality TFBS predictions. We believe that UniBind provides a critical resource that will enable an array of studies aiming at better understanding the underlying mechanisms of transcriptional regulation.

References

1. Rohs, R. et al. *Nature* **461**, 1248–1253 (2009).

2. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. *Science* **316**, 1497–1502 (2007).

3. Teytelman, L., Thurtle, D.M., Rine, J. & Oudenaarden, A. van *Proc. Natl. Acad. Sci.* **110**, 18602–18607 (2013).

4. Jain, D., Baldi, S., Zabel, A., Straub, T. & Becker, P.B. *Nucleic Acids Res.* **43**, 6959–6968 (2015).

5. Wreczycka, K., Franke, V., Uyar, B., Wurmus, R. & Akalin, A. *bioRxiv* 107680 (2017).doi:10.1101/107680

6. Hunt, R.W. & Wasserman, W.W. *Genome Biol.* **15**, 412 (2014).

7. Mathelier, A. & Wasserman, W.W. *PLoS Comput. Biol.* **9**, e1003214 (2013).

8. Mathelier, A. et al. *Cell Syst.* 1–9 (2016).

9. Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. *Nucleic Acids Res.* **46**, D267–D275 (2018).

10. Khan, A. et al. *Nucleic Acids Res.* **46**, D260–D266 (2018).

11. Hunt, R.W., Mathelier, A., Peso, L. del & Wasserman, W.W. *BMC Genomics* **15**, 472 (2014).

12. Patel, R.Y. & Stormo, G.D. *Bioinformatics* **30**, 941–948 (2014).

13. Kapur, J.N., Sahoo, P.K. & Wong, A.K.C. *Comput. Vis. Graph. Image Process.* **29**, 273–285 (1985).

14. Newburger, D.E. & Bulyk, M.L. *Nucleic Acids Res.* **37**, D77–D82 (2009).

15. Yamada, N., Lai, W.K., Farrell, N., Pugh, B.F. & Mahony, S. *bioRxiv* 266536 (2018).doi:10.1101/266536

16. The ENCODE Project Consortium *Science* **306**, 636–640 (2004).

17. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. *Genome Biol.* **9**, S4 (2008).
18. Tonekaboni, S.A.M., Mazrooei, P., Kofia, V., Haibe-Kains, B. & Lupien, M. *bioRxiv* 222562 (2018).doi:10.1101/222562

Keywords: TF, DNA ineractions, TFBS, ChIP, seq

Multilevel logical modelling of the regulatory network governing dorsal-ventral axis specification in the sea urchin P. lividus.

Swann Floc'hlay * [1], Céline Hernandez [2], Morgane Thomas-Chollier [3], Thierry Lepage [4], Denis Thieffry [5]

[1] Institute of Biology at the Ecole Normale Superieure (IBENS) – CNRS : UMR8197 – France
[2] Institut de biologie de l'ENS (IBENS) – CNRS : UMR8197 – France
[3] Institute of Biology at the Ecole Normale Superieure (IBENS) – CNRS : UMR8197 – 46 rue d'Ulm 75005 Paris, France
[4] Institut de Biologie Valrose (IBV) – CNRS : UMR7277 – France
[5] Institut de biologie de l'ENS (IBENS) – CNRS : UMR8197 – France

Located at the basis of the deuterostome branch, echinoderms occupy a unique position to study the regulatory networks governing embryo morphogenesis. The dorsal-ventral (D-V) axis specification in the sea urchin *Paracentrotus lividus* is controlled by various transcription factors, including two TGF-βs: Nodal and BMP2/4. However, the signalling network downstream of these key morphogens is not yet fully understood. To identify Nodal and BMP2/4 target genes, we have performed a systematic functional analysis using RNA sequencing and in situ hybridization screens [1-4]. The analysis of these data enables to delineate various novel interactions.
To gain further insights into this developmental process, we have developed a predictive dynamical model of the corresponding signalling/regulatory network. More specifically, using a logical modelling framework implemented in the software GINsim [5], we account for the specification of three main ectodermal regions along the D-V axis (ventral, ciliary and dorsal ectoderm) in terms of specific marker gene expression patterns

In our model analysis, we first focused on the computation of stable states and on their reachability in single representative cells, depending on signalling inputs. These model simulations correctly reproduce wild-type and mutant phenotypes. They also unravel complex mechanisms linked to admp1 inhibition during Nodal over-expression. Our model further account for the role of Panda in Nodal inhibition.

Finally, taking advantage of the software EpiLog [6], we have simulated grids of cells connected through signalling molecules, thereby reproducing various reported patterns.

References

*Speaker

1. Duboc V, Lapraz F, Saudemont A, Bessodes N, Mekpoh F, Haillot E, Quirin M, Lepage T (2010). Nodal and BMP2/4 pattern the mesoderm and endoderm during development of the sea urchin embryo. Development 137(2): 223-35.

2. Saudemont A, Haillot E, Mekpoh F, Bessodes N, Quirin M, Lapraz F, Duboc V, Ŕottinger E, Range R, Oisel A, Besnardeau L, Wincker P, Lepage T (2010). Ancestral regulatory circuits governing ectoderm patterning downstream of Nodal and BMP2/4 revealed by gene regulatory network analysis in an echinoderm. PLoS Genet 6(12): e1001259.

3. Molina MD, de Crozé N, Haillot E, Lepage T (2013). Nodal: master and commander of the dorsal-ventral and left-right axes in the sea urchin embryo. Curr Opin Genet Dev 23(4): 445-53.

4. Lapraz F, Haillot E, Lepage T (2015). A deuterostome origin of the Spemann organiser suggested by Nodal and ADMPs functions in Echinoderms. Nat Commun 6: 8434.

5. http://ginsim.org/

6. http://epilog-tool.org/

Keywords: Computational systems biology, embryonic development, dynamical modelling, gene networks

ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments

Jeanne Chèneby *† [1,2], Marie Artufel [1,2], Marius Gheorghe [3], Anthony Mathelier [4,5], Benoit Ballester‡ [1,2]

[1] INSERM, UMR1090 TAGC, Marseille, France – Centre de Recherche Inserm – France
[2] Aix-Marseille Université, UMR1090 TAGC, Marseille, France – Aix-Marseille Université - AMU – France
[3] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo (NCMM) – Norway
[4] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway – Norway
[5] Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway – Norway

Highlight abstract:
Context

Regulation of transcription plays an important role in the establishment and maintenance of structure and function of cells, tissues and ultimately the whole eukaryotic organism.

Transcription Regulators (TR) are composed of transcription factors (TFs), transcriptional coactivators (TCAs) and chromatin-remodeling factors (CRFs), they drive gene transcription and the organization of chromatin through DNA binding. For a variety of DNA-binding proteins a genome-wide map of binding site across many cell types and tissues have been obtained thanks to the development high throughput sequencing technique such as chromatin immunoprecipitation and sequencing (ChIP-seq). Hundreds of TR occupancy maps are available with the rapid accumulation of ChIP-seq data in public databases. Integrative studies would offer significant insights into the dynamic mechanisms by which a TF selects its binding regions in each cellular environment.

The integration processes and the underlying detection of TF binding regions are challenging because of the heterogeneity of the pipelines used to process the numerous raw ChIP-seq data, as well as the variety of underlying formats used. Integration therefore require first meta data curation and a uniform reprocessing of the raw ChIP-seq data.

ReMap 2015 [2] has been the first large scale integrative initiative in which raw ChIP-seq data were uniformly reprocessed and quality controlled, offering significant insights into the com-

*Speaker

†Corresponding author: jeanne.cheneby@inserm.fr

‡Corresponding author: benoit.ballester@inserm.fr

plexity of the human regulatory landscape. This work present the latest update of the ReMap database leading to the largest regulatory catalogue in human.

Methodology

The crucial point in ChIP-seq data integration is the standardisation and curation of experimental annotation since no guidelines are proposed to submit ChIP-seq experiments in public warehouse such as Gene Expression Omnibus[3] (GEO) and ArrayExpress[4] (AE). In this ReMap update we manually curated 2,349 ChIP-seq experiments from GEO and AE. Those 2,349 experiments were consolidated with 1160 datasets from Encyclopedia of DNA Elements[5] (ENCODE) allowing the processing of 3,500 ChIP-seq experiments for 485 TRs and 346 cell lines/tissues.

Those standardized ChIP-seq where then processed using our uniform analysis pipeline updated from the ReMap 2015 work.

A key step in the ReMap integrative project is the control quality procedures which consist in data filtering using three different metrics assessing the signal-to-noise ratios (normalized strand coefficient, NSC and relative strand correlation, RSC) and the specificity of the immunoprecipitation (fraction of reads in peaks, FRiP). Among the 3,500 ChIP-seq datasets, 2,829 pass the quality criterions, thus forming the core datasets of the ReMap 2018 catalogue. This catalogue consists in an large file containing 80 million genomic position of binding regions of 485 TRs in 346 cell lines and tissues. In addition, we developed a workflow to create genomic position of binding regions for a specific TR and/or cell lines or tissues.

This work is accessible via a website allowing for browsing and downloading the catalogues (http://remap.cisreg.eu). We propose in the ReMap 2018 website a dynamic search of TRs, aliases, cell lines and experiments. We also improved the capacity to download all ReMap files in bulk or individually.

Finally, to enhance accessibility of our data to end users we made ReMap data available in all UCSC[6] and Ensembl Genome Browsers[7] mirror sites by providing public sessions and Track Hubs.

Conclusion

In 2015, the ReMap database was introduced to capture the genome regulatory space by integrating public ChIP-seq datasets, covering 237 TRs across 13 million binding regions. In the ReMap 2018 update, this catalog was largely extended to constitute a unique collection of regulatory regions with 80 million binding regions. Specifically, we have processed 3,500 ChIP-seq datasets and after analysis and quality control we retained a total of 2,829 high confident ChIP-seq datasets covering a total of 485 TRs in a catalog of 80 million binding regions.

Ref:

Jeanne Chèneby, Marius Gheorghe, Marie Artufel, Anthony Mathelier, Benoit Ballester; ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments, Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018

Aurélien Griffon, Quentin Barbier, Jordi Dalino, Jacques van Helden, Salvatore Spicuglia, Benoit Ballester; Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape, Nucleic Acids Research, Volume 43, Issue 4, 27 February 2015

Barrett T.,Wilhite S.E.,Ledoux P.,EvangelistaC.,Kim I.F.,Tomashevsky M.,Marshall K.A.,

Phillippy K.H.,Sherman P.M.,Holko M.et al. NCBI GEO: archive for functional genomics data sets–update.Nucleic Acids Res.2013.

Kolesnikov N., Hastings E., Keays M., Melnichuk O., Tang Y.A., Williams E., Dylag M.,Kurbatova N., Brandizi M., Burdett T. et al. ArrayExpress update–simplifying data submissions. Nucleic Acids Res. 2015

ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. Nature . 2012

Tyner C.,Barber G.P.,Casper J.,Clawson H.,Diekhans M.,Eisenhart C.,Fischer C.M.,Gibson D.,Gonzalez J.N.,Guruvadoo L. et al. The UCSC Genome Browser database: 2017 update.

Nucleic Acids Res. 2017

Aken B.L.,Achuthan P.,Akanni W.,Amode M.R.,Bernsdorff F.,Bhai J.,Billis K.,Carvalho-Silva D. ,Cummins C.,Clapham P. et al. Ensembl 2017. Nucleic Acids Res. 2017

Keywords: Transcription Fractor, Integration, Database, DNA, binding

Logical modelling and analysis of cellular regulatory networks with GINsim 3.0

Aurélien Naldi * [1], Pedro Monteiro [2], Céline Hernandez [1], Wassim Abou-Jaoudé [1], Claudine Chaouiya [3], Denis Thieffry * † [1]

[1] Institut de Biologie de l'ENS (IBENS) – Institut National de la Santé et de la Recherche Médicale - INSERM : U1024, Centre National de la Recherche Scientifique - CNRS : UMR8197 – 46 rue d'Ulm, 75005 Paris, France

[2] INESC-ID/Instituto Superior Técnico – Rua Alves Redol 9, P-1000-029 Lisboa, Portugal

[3] Instituto Gulbenkian de Ciência (IGC) – Apartado 14 P-2781-901 Oeiras, Portugal

The logical formalism is a well established qualitative modeling framework to study large cellular networks with scarce kinetic data [1]. A logical model is composed of a set of components, each associated to an activity level. Boolean components can be either active or inactive, while multi-valued ones are assigned additional activity levels. Signed interactions between components denote regulatory mechanisms. Finally, logical rules associated with the components further define the combined effects of their regulators.

The "simulation" of the resulting logical model generates a graph representing dynamical behaviours starting from a (set of) initial state(s) and applying a specific updating scheme (e.g. synchronous or asynchronous). As this step can be intractable for large models, dedicated analysis methods have been designed to identify some dynamical properties more efficiently, including model reduction and graph compression techniques, as well as algorithms to compute stable states and other kinds of attractors.

Reaching its fifteenth anniversary, the GINsim software suite provides an interactive graphical user interface for the definition and analysis of logical models [2]. The last official version of GINsim (release 3.0) includes various novel tools, including the computation of local regulatory graphs, new state updating schedules, as well as the possibility to perturb interactions. Furthermore, this release includes novel export functionalities, including SBML qual, as well as MaBoSS and Pint input files. The use of GINsim 3.0 will be illustrated through the definition, the analysis and the simulation of a small, yet sophisticate, model of the mammalian p53-Mdm2 network.

References

1. Abou-Jaoudé W, Traynard P, Monteiro P, Saez-Rodriguez J, Helikar T, Thieffry D, Chaouiya D (2016). Logical modeling and dynamical analysis of cellular networks. Frontiers in Genetics 7: 94.

2. http://ginsim.org

[*] Speaker

[†] Corresponding author: thieffry@ens.fr

Keywords: Logical modeling, regulatory networks

PhyloSofS: PHYLOgenies of Splicing isOForms Structures

Adel Ait-Hamlat [1], Lélia Polit [1], Diego Javier Zea * [1], Hugues Richard† [1], élodie Laine‡ [1]

[1] Biologie Computationnelle et Quantitative = Laboratory of Computational and Quantitative Biology (LCQB) – Université Pierre et Marie Curie - Paris 6, Centre National de la Recherche Scientifique : UMR7238 – Biologie Computationnelle et Quantitative UMR 7238 CNRS-Université Pierre et Marie Curie Site des Cordeliers Bât. A - 4ème étage, 15, Rue de lÉcole de Médecine 75006 Paris, France, France

Introduction

Alternative splicing (AS) has the potential to greatly expand the proteome in eukaryotes, by producing several transcript isoforms from the same gene. It has been associated with multiple biological functions, such as regulation of intermolecular interactions and developmental programmes (Kelemen et al. 2013; Bush et al. 2017). AS deregulation has been associated with the development of various diseases, particularly with neurodegenerative disorders and cancer (Ward and Cooper 2009). Although AS is well described at the genomic level, little is known about its contribution to protein evolution and its impact on protein structure.

Here we present PhyloSofS, a fully automated computational tool that infers plausible evolutionary scenarios explaining a set of transcripts observed in several species and models the three-dimensional structures of the produced isoforms. The method is applicable at large scale and provides a mean to address unresolved questions linked to AS. First, the structural models of the different protein isoforms can help us identify alternative splicing events (ASEs) inducing substantial conformational rearrangements or even fold changes and discovering new therapeutic targets (isoforms). More generally, it shall also shed light on the evolutionary paths leading to functional innovation.

As a first case study, we applied PhyloSofS to the c-Jun N-terminal kinase (JNK) family. JNKs bear a great interest for medicinal research as they are involved in crucial signalling pathways and some of their related disorders were associated to particular JNK isoforms (Brecht et al. 2005). This family also represents a high degree of complexity, with 60 observed transcripts composed by a total of 19 different exons, most of the transcripts comprising more than 10 exons, and high disparities between species, from 1 to 8 transcripts per gene per species.

In a second step, and to bring the analysis to a larger scale, we are now engaging on the study of a dozen genes and gene families where different protein functions have been linked to AS experimentally. As future work, we plan to extend the analysis to the proteome level to increase our knowledge of protein isoform evolution and structure.

Materials and methods

*Speaker

†Corresponding author: hugues.richard@upmc.fr

‡Corresponding author: elodie.laine@upmc.fr

PhyloSofS performs two main tasks, the phylogenetic reconstruction of the transcript history and the molecular modelling of the isoforms. The phylogenetic reconstruction algorithm takes a binary gene tree for a set of species and their ensemble of transcripts as an input. A forest of phylogenetic trees describing plausible evolutionary scenarios that can explain the observed transcripts is reconstructed using the maximum parsimony principle. The generation of structural models for each isoform is performed using comparative modelling. PhyloSofS uses the HH-suite (Hildebrand et al. 2009) to look for homologous structures and MODELLER (Martí-Renom et al. 2000) to perform the homology modelling step. The generated models are automatically checked using PROCHECK to assess their quality (Laskowski, MacArthur, and Thornton 2012). The user can further evaluate and compare the stability and dynamical behaviour of these models using molecular dynamics.

For studying the JNK family, the peptide sequences of all splice variants observed in human, mouse, xenopus, zebrafish, fugu, drosophila, and nematode were retrieved from Ensembl (Zerbino et al. 2018) along with the phylogenetic gene tree. The homologous exons were identified by aligning the sequences with MAFFT (Katoh and Standley 2013) and projecting the alignment on the human annotation. PhyloSofS was used to reconstruct the phylogenetic forest using one million iterations to retain the most parsimonious evolutionary scenario. PhyloSofS was also used to generate 3D molecular models for the corresponding protein isoforms. We subsequently performed molecular dynamics simulations of 3 human isoforms, starting from the predicted structural models.

Results

The forest reconstructed by PhyloSofS for JNK is comprised of 7 transcript trees. The root of each tree corresponds to the appearance of its transcript in evolution, indicating where a new ASE occurred. We observed some transcript losses and 11 mutations (inclusion or exclusion of exons) along the JNK transcripts' phylogeny. We also observed 14 orphan leaves that correspond to transcripts for which no phylogeny could be reconstructed. These transcripts are not conserved across the studied species, and thus are likely non-functional.

The analysis of the transcripts' phylogeny inferred by PhyloSofS for JNKs emphasized several characteristics of the evolution of this protein family. First, it revealed a rather low number of mutations, illustrating the fact that the sequences of the JNK genes and their exon sites are highly conserved through evolution. Second, it enabled to date ASEs associated to two pairs of mutually exclusive exons. One of them is of particular interest because the two exons are homologous and were shown to modulate the affinity of JNKs to their cellular substrates. Key residues responsible for such modulation were identified by characterizing their flexibility in solution with molecular dynamics simulations using the structural models generated by PhyloSofS. Our phylogenetic reconstruction revealed that the most recent common ancestor of all 7 species contained only one transcript with only one of this two homologous exons. The other exon appeared in the ancestor common to mammals, amphibians and fishes. Our results suggest that it is issued from the duplication of the more ancient exon before the duplication of the ancestral JNK gene. Our analysis also highlighted 2 transcripts specific to JNK3 across several species and showed that the duplicated exon is not expressed in the JNK3 sub-forest. This may suggest a subfunctionalization for JNK3, which is the only paralogue being specifically expressed in certain tissues, namely the heart, brain and testes. Finally, our results also highlighted an ASE inducing a large deletion, yet conserved across several species. The resulting isoform is stable in solution and could play a role in the cell.

Conclusion

PhyloSofS allows to obtain information about the evolutionary conservation and the structural

stability of protein isoforms. It has been proven in our case study that PhyloSofS helps identifying functional ASEs that can be further studied. The phylogenetic analysis with PhyloSofS is also able to date the appearance of protein isoforms, which can help in the understanding of their biological functions and evolutionary history. Also, it can be used to simply cluster together related transcripts that have jointly undergone inclusion/exclusion of some exons.
Based on this encouraging preliminary results, we are now applying PhyloSofS to a curated set of a dozen proteins and protein families that possess at least one ASE with known biological implication. Some of these proteins present new challenges, such as, for example, few homologues with known structures, very large gene sequences, or uneven species representations. All those peculiarities help us to further improve and generalize the approaches used in PhyloSofS. One of the improvements which we are working on is to complement gene models annotation using curated RNA-Seq and Ribo-Seq evidence across multiple tissues and species using the Bgee database (Bastian et al., n.d.).

References Bastian, Frederic, Gilles Parmentier, Julien Roux, Sebastien Moretti, Vincent Laudet, and Marc Robinson-Rechavi. n.d. ”Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species.” In Lecture Notes in Computer Science, 124–31.
Brecht, Stephan, Rainer Kirchhof, Ansgar Chromik, Mette Willesen, Thomas Nicolaus, Gennadij Raivich, Jan Wessig, et al. 2005. ”Specific Pathophysiological Functions of JNK Isoforms in the Brain.” The European Journal of Neuroscience 21 (2): 363–77.
Bush, Stephen J., Lu Chen, Jaime M. Tovar-Corona, and Araxi O. Urrutia. 2017. ”Alternative Splicing and the Evolution of Phenotypic Novelty.” Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 372 (1713). https://doi.org/10.1098/rstb.2015.0474.
Hildebrand, Andrea, Michael Remmert, Andreas Biegert, and Johannes S´oding. 2009. ”Fast and Accurate Automatic Structure Prediction with HHpred.” Proteins: Structure, Function, and Bioinformatics 77 (S9): 128–32.
Katoh, Kazutaka, and Daron M. Standley. 2013. ”MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” Molecular Biology and Evolution 30 (4): 772–80.
Kelemen, Olga, Paolo Convertini, Zhaiyi Zhang, Yuan Wen, Manli Shen, Marina Falaleeva, and Stefan Stamm. 2013. ”Function of Alternative Splicing.” Gene 514 (1): 1–30.
Laskowski, R. A., M. W. MacArthur, and J. M. Thornton. 2012. ”PROCHECK: Validation of Protein-Structure Coordinates.” In International Tables for Crystallography, 684–87.
Martí-Renom, M. A., A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Sali. 2000. ”Comparative Protein Structure Modeling of Genes and Genomes.” Annual Review of Biophysics and Biomolecular Structure 29: 291–325.
Ward, Amanda J., and Thomas A. Cooper. 2009. ”The Pathobiology of Splicing.” The Journal of Pathology. https://doi.org/10.1002/path.2649.
Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. ”Ensembl 2018.” Nucleic Acids Research 46 (D1): D754–61.

Keywords: alternative splicing, transcript phylogenies, molecular modelling, isoform evolution

Prediction of new multiciliogenesis genes using a fine-grained comparative genomic approach

Audrey Defosset *† [1], Yannis Nevers [1], Arnaud Kress [1], Raymond Ripp [1], Olivier Poch [1], Odile Lecompte‡ [1]

[1] Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (ICube) – ENGEES, Institut National des Sciences Appliquées [INSA] - Strasbourg, université de Strasbourg, CNRS : UMR7357 – 300 bd Sébastien Brant - BP 10413 - F-67412 Illkirch Cedex, France

Context

Cilia are small microtubule-based organelles protruding from the surface of a wide array of eukaryotic cells. They are involved in various functions (locomotion, fluid flow, perception of the environment, development, etc.), depending on cell type, species and development stage. Most vertebrate cells possess the ability to generate single non-motile cilia, known as primary cilia. These structures can serve as sensory devices capable of transducing signals during development and homeostasis. Certain specialized cells, however, are able to assemble up to three hundred motile cilia that beat coordinately in order to produce a directional fluid flow. In humans, these cells are mainly found in the respiratory tract, where they help with mucus clearance, as well as in the ventricular system of the brain, where they are essential for the circulation of cerebrospinal fluid, and in both female and male reproductive tracts. Dysfunctions of these cells are associated with a variety of genetic diseases, called ciliopathies, including primary ciliary dyskinesia, with a very diverse range of phenotypes. Symptoms usually associated with defects in multiciliation include an increased risk of hydrocephalus, chronic respiratory infections, as well as sterility. Despite the fundamental and biomedical importance of multiciliogenesis, the molecular and cellular mechanisms involved in the process are still poorly understood. It is therefore important, both for the field of biology and for medical diagnostics and therapeutics, to identify the different actors involved in multiciliogenesis, and understand their interactions.

In recent years, comparative genomics has been established as a classical 'in silico' approach to study specific processes, understand their evolution and determine which genes are involved in their mechanisms. A typical approach is phylogenetic profiling, for which it is essential to have knowledge of any loss or abnormality of the process of interest in subsets of species. It is admitted that genes participating in the same mechanism will generally be conserved together in the species that possess it, in order to maintain the integrity of the process. In contrast, they will generally be lost together in species lacking this process. Thus, genes involved in the same process often present similar taxonomic distributions. This genotype/phenotype correlation can be exploited to predict which genes are involved in a process by studying their taxonomic distribution (Pellegrini et al., 1999). It has been successfully applied to various processes, including

*Speaker

†Corresponding author: adefosset@etu.unistra.fr

‡Corresponding author: odile.lecompte@unistra.fr

cilia, which are ancestral organelles present in the last eukaryotic common ancestor, and that exhibit a peculiar evolutionary history, with various independent losses in the eukaryotic lineage (Li et al., 2004; Nevers et al., 2017).

Results and discussion

Little information has been compiled on the distribution of multiciliation in eukaryotes. It has been observed in most Metazoa, in the spermatozoids of some plants, and in unicellular organisms, but so far, no similarity has been reported between the mechanisms existing in the different eukaryotic Kingdoms. Concerning the Metazoa, knowledge about the presence and absence of multiciliation remains sparse outside of Vertebrates. It seems to be absent in Ecdysozoa (Arthropodes and Nematodes), and a specific group of fish, the Otomorpha (*Danio rerio, Astyanax mexicanus...*), appears to have an incomplete multiciliogenesis with a reduced number of cilia on their multiciliated cells. As a first step in our effort to identify new multiciliated genes, we defined the precise taxonomic distribution of the current key genes of multiciliogenesis in Metazoa. In this evolutionary study, we highlighted an absence of most genes, such as CEP63, CCDC78 and CEP152, in Ecdysozoa, which correlates with the apparent absence of multiciliogenesis in this clade, as well as a loss of CDC20B in Otomorpha. Using phylogenetic profiling and the OrthoInspector orthology prediction program (Linard et al., 2015), we searched for human genes conserved in most Metazoa but lost in Ecdysozoa and Otomorpha. This preliminary search resulted in a highly unspecific set of genes, due maintly to the fact that multiciliation is not the only process that was lost in Ecdysozoa. It also appeared that, to be effective, phylogenetic profiling based on presence/absence of genes has to take into account very atypical evolutionary histories, with complete losses in various species and taxa, and that a subtler approach is required in the case of multiciliation.

We capitalized on the incomplete multiciliogenesis observed in the Otomorpha group of fish to develop a new strategy. A detailed analysis of multiple alignments of protein families known to be involved in multiciliation revealed several specificities in the Otomorpha group of fish, notably absent or divergent regions in the protein sequences of CCNO, involved in the formation and migration of centrioles, and MCIDAS, the central activator of multiciliogenesis, capable of regulating the transcription of various genes. We surmised that this was related to the seemingly incomplete multiciliogenesis that has been observed in these species. In lights of these results, we hypothesized that total absence of genes in Otomorpha might be a too strict criterion to explain the incomplete nature of multiciliation. Therefore, we searched for other genes presenting abnormal divergence in these species. Current methods used to detect protein region divergence or loss are mostly based on evolutionary distance between sequences, either between two species, such as the Reciprocal Smallest Distance (Wall et al., 2003), or between genomic sequences in a multiple sequence alignment (Kumar et al., 2016). While these are effective methods for specific evolutionary questions, they are not suitable for large scale searches involving complete proteomes and a large number of species. Our current problematic required a fast approach capable of detecting sequence conservation abnormalities for a specific subset of species in the complete human proteome.

To address this, we developed a new approach to identify, on a whole proteome level, proteins that present an atypical pattern of conservation among different species resulting from divergence or loss of a domain and/or motif in a specific taxon. Such clade-specific divergence can lead to unusual ranking of species in BLAST similarity searches. Generally, the more recent the separation between species is, the more similar their protein sequences will be, and as such, it is possible to assess taxonomic proximities through sequence similarity. Our method is based on the hypothesis that in a "classic" case, the succession of hits in a BLAST result will approximately respect a defined taxonomic order, whereas for proteins presenting a differential conservation, the order will be altered and two usually close taxa will have different ranks. The program we developed, called BLUR (Blast Unexpected Ranking), analyses the global rankings of two related taxa in all BLASTP outputs of a complete proteome, as well as the similarity between the sequences of both groups by comparing bitscore and E-value ratios and average distances to the query sequence, in order to detect cases of abnormal divergence and differential conservation. A statistical analysis is then performed on the distribution of each of these criteria, using the Tukey's fences statistical method to identify outliers and to classify proteins into high, mid or low priority, depending on how many of the criteria were detected as outliers.

We applied this concept to the human proteome by comparing sequence conservation of human proteins in two groups of fish: the Otomorpha (exhibiting incomplete multiciliation) and the Euteleosteomorpha (with complete multiciliation). The protein database used to perform the BLASTP searches contained 735 complete proteomes, including 9 Otomorpha and 13 Euteleosteomorpha. The processing of the BLAST results with BLUR generated a list of 167 high priority, 718 mid priority and 2057 low priority proteins. So far, 87 of these proteins have been confirmed as being either absent or divergent in Otomorpha by manual inspection of multiple sequence alignments. Among them, 3 are functionally related to the WNT-pathway, 5 code for homeobox or homeodomain-interacting proteins, and 18 are microtubule or centrosome related genes. In the next step, these promising results will be enriched by the integration of further evolutionary data, as well as functional and medical data from various sources, such as results from transcriptomics experiments or known variants and pathologies linked to multiciliation defects. Based on these data, the target genes will be prioritized and the most promising will be experimentally validated. All these results will then be integrated in a knowledge base on multiciliation with the purpose of finely characterizing and predicting genes and networks involved in this process.

Conclusions and perspectives

We have developed a tool capable of rapidly detecting differential conservation from a BLAST search result on the whole proteome level, regardless of the underlying biological process. Our goal was to provide users with the ability to detect evolutionary divergences that go beyond mere presence/absence of genes. After further development, BLUR will be made available online, with precomputed BLAST searches for *Homo sapiens* and model organisms representing some of the major life phyla, such as *Mus musculus*, *Drosophila melanogaster*, *Arabidopsis thaliana*, or *Saccharomyces cerevisiae*. The user will then be able to select any taxa of interest in all life domains to be studied, while retaining an effective detection power. Furthermore, atypical conservation patterns detected by BLUR will be automatically analyzed with PROBE (Kress et al., in press), a tool that identifies conserved protein blocks in multiple sequence alignments.

Kress, A., Lecompte, O., Poch, O., Thompson, J.D. PROBE: analysis and visualization of protein block-level evolution. Bioinformatics (in press)

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol *33*, 1870–1874.

Li, J.B., Gerdes, J.M., Haycraft, C.J., Fan, Y., Teslovich, T.M., May-Simera, H., Li, H., Blacque, O.E., Li, L., Leitch, C.C., et al. (2004). Comparative Genomics Identifies a Flagellar and Basal Body Proteome that Includes the BBS5 Human Disease Gene. Cell *117*, 541–552.

Linard, B., Allot, A., Schneider, R., Morel, C., Ripp, R., Bigler, M., Thompson, J.D., Poch, O., and Lecompte, O. (2015). OrthoInspector 2.0: Software and database updates. Bioinformatics *31*, 447–448.

Nevers, Y., Prasad, M.K., Poidevin, L., Chennen, K., Allot, A., Kress, A., Ripp, R., Thompson, J.D., Dollfus, H., Poch, O., et al. (2017). Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling. Mol Biol Evol *34*, 2016–2034.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A *96*, 4285–4288.

Wall, D.P., Fraser, H.B., and Hirsh, A.E. (2003). Detecting putative orthologs. Bioinformatics *19*, 1710–1711.

Keywords: comparative genomics, cilia, multiciliation, evolution, phylogeny, genetic diseases

RSAT 2018: regulatory sequence analysis tools 20th anniversary

Nga Thi Thuy Nguyen [1], Bruno Contreras-Moreira [2], Jaime Castro-Mondragon [3], Walter Santana-Garcia [4], Raul Ossio [4], Carla Daniela Robles-Espinoza [4], Mathieu Bahin [1], Samuel Collombet [1], Pierre Vincens [1], Denis Thieffry [1], Jacques Van Helden [5], Alejandra Medina Rivera [4], Morgane Thomas-Chollier *† [1]

[1] Institut de Biologie de l'Ecole Normale Superieure (IBENS) – INSERM 1024 CNRS 8197 – 46 rue d'Ulm 75005 Paris, France
[2] Estacion Experimental de Aula Dei-CSIC – Zaragoza, Spain
[3] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway (NCMM) – Norway
[4] Laboratorio Internacional de Investigacion sobre el Genoma Humano (LIIG) – Universidad Nacional Autonoma de Mexico, Campus Juriquilla, Blvd Juriquilla 3001, Santiago de Quere taro 76230, Mexico
[5] Theory and Approaches of Genome Complexity (TAGC) – Institut National de la Santé et de la Recherche Médicale - INSERM : UMRS 1090, Aix-Marseille Université - AMU – Marseille, France

RSAT (Regulatory Sequence Analysis Tools) is a suite of modular tools for the detection and the analysis of cis-regulatory elements in genome sequences. Its main applications are (i) motif discovery, including from genome-wide datasets like ChIP-seq/ATAC- seq, (ii) motif scanning, (iii) motif analysis (quality assessment, comparisons and clustering), (iv) analysis of regulatory variations, (v) comparative genomics. Six public servers jointly support 10 000 genomes from all kingdoms. The latest novel or refactored programs include updated programs to analyse regulatory variants (retrieve-variation-seq, variation-scan, convert-variations), along with tools to extract sequences from a list of coordinates (retrieve-seq-bed), to select motifs from motif collections (retrieve-matrix), and to extract orthologs based on Ensembl Compara (get-orthologs-compara). This Anniversary update gives a 20-year perspective on the software suite. RSAT is well-documented and available through Web sites, SOAP/WSDL (Simple Object Access Protocol/Web Services Description Language) web services, virtual machines and stand-alone programs at http: //www.rsat.eu/.

The presentation will provide an overview of the suite, highlighting the most recent developments, and illustrated with a use-case covering the tools matrix-clustering and retrieve-matrix.

The article has been accepted in the Nucleic Acids Research Web Server Issue 2018 :

Nguyen NTT, Contreras-Moreira B et al, Nucleic Acids Research, 2018 in press

https://doi.org/10.1093/nar/gky317

*Speaker

†Corresponding author: mthomas@biologie.ens.fr

Keywords: cis, regulatory elements, motif analysis, motif discovery, regulatory variants

Predicting 3'UTR's regulation of protein multifunctionality

Diogo Ribeiro *† [1], Adrien Teixeira [1], Andreas Zanzoni [1], Lionel Spinelli [1], Christine Brun [1]

[1] Theories and Approaches of Genomic Complexity (TAGC) – Aix Marseille Université, Institut National de la Santé et de la Recherche Médicale : U1090, Centre National de la Recherche Scientifique – Parc scientifique de Luminy, 163 avenue de Luminy, 13288 Marseille cedex 9, France

Characteristics of extreme protein multifunctionality

Constructing a complex organism does not require a large number of genes. Rather, organism complexity is provided by the ensemble of all available functions and their regulation. **Protein multifunctionality, like alternative splicing, allows cells to make more with less**. A canonical example of an extreme multifunctional (and moonlighting) protein is the human aconitase, an enzyme of the tricarboxylic acid cycle (TCA cycle) that also functions as a translation regulator, upon a conformational change. Often, the extreme multifunctionality of proteins is also brought upon the gain of new interactors, its presence on a different tissue or a change in its cellular localisation.

In order to better understand globally the functional plasticity of proteins, **we have computationally identified 238 human "extreme multifunctional proteins" (EMFs)** using a in-house approach that utilizes protein-interaction networks and protein annotations (Chapple et al., 2015). These predictions, as well as other manually curated multifunctional proteins, were recently made available in MoonDB (http://moondb.hb.univ-amu.fr/).

EMFs possess characteristics that set them apart from other proteins. Within a protein interactome, a typical EMF is likely to have a high number of protein partners and to be central to the network. Importantly, EMFs are more likely to be involved in multiple diseases (Zanzoni, Chapple and Brun, 2015) and to be expressed ubiquitously, suggesting that they can perform alternative functions in different tissues (Chapple et al., 2015). In addition to be expressed in more tissues, analysing cellular component GO term annotations, we have recently found that **EMFs are prone to localise to 'unlikely' and distant combinations of cellular locations**, such as nucleus and membrane or mitochondria and extracellular matrix. Moreover, EMFs contain more short linear motifs (SLiMs) than other proteins. These shorts conserved sequences are mostly located in structurally disordered regions and are involved in transient interactions and mediate molecular switches between functions. These results provide insights on how the changement of function of these multifunctional proteins may occur.

At the sequence level, using the APADB and PolyAsite databases, we found that mRNAs encoding EMFs have longer 3' untranslated regions (3'UTRs) and bear a higher number of al-

*Speaker

†Corresponding author: diogo.ribeiro@inserm.fr

ternative polyadenylation sites than mRNAs encoding other proteins. Correspondingly, using Ensembl transcript models, we found that **EMFs are annotated in with more alternative 3'UTRs than most proteins.** This suggests that genes encoding EMFs may be regulated by elements located in their 3'UTRs.

3'UTR regulation of subcellular localisation via protein complex formation

Interestingly, usage of alternative 3'UTRs has recently been found to influence the functional fate of its cognate proteins (Berkovits and Mayr, 2015). By recruiting RNA-binding proteins (RBPs) to the site of translation, **3'UTRs were shown to affect the function of its cognate proteins by promoting the co-translational formation of protein complexes** that interact with the nascent peptide chain (Berkovits and Mayr, 2015; Mayr, 2016, 2017).

The relationship between alternative 3'UTRs, subcellular localization and protein complex formation has been recently demonstrated in details by Berkovits & Mayr (Berkovits and Mayr, 2015) for CD47, a ubiquitous protein involved in a range of cellular processes, including apoptosis, adhesion, migration and preventing phagocytosis by macrophages. Whereas the CD47 protein translated from a short 3'UTR-mRNA is retained in the Endoplasmic Reticulum (ER), the protein translated from the long 3'UTR-mRNA localizes to the plasma membrane (PM). This contrasting cell distribution is achieved through the recruitment by the long 3'UTR-mRNA of specific protein partners necessary for addressing CD47 to the PM. Formation of this sorting complex is mediated by a RBP (ELAVL1) recognizing a binding site on the long 3'-UTR and absent from the short one.

These complexes thus involve the following components: 1) an mRNA with a 3'UTR; 2) the cognate protein being translated (hereby known as nascent protein; 3) an RBP able to bind the 3'UTR; 4) an effector protein (hereby intermediate protein), recruited by the RBP and/or the nascent protein and predicted to alter the function of the nascent protein (e.g. by transportation to a new cellular location). **As this mechanism has been described only for CD47 and inferred for few other cases, there is a need to interrogate its full prevalence** in the cell and determine whether the use of alternative 3'UTRs is a major contributor to the diversification of protein function.

Given the high propensity of EMF proteins to use alternative 3'UTRs and to be localised in distinct cellular components, these proteins constitute a very pertinent model system in which to study the role of 3'UTRs in regulating protein function. Hence**, we set out to reveal the extent of the participation of 3'UTRs to protein complex formation in human EMF proteins**, and to understand how this could affect protein multifunctionality. To this aim, we predicted an extensive list of all available 3'UTR-protein complexes, based on protein-protein and protein-mRNA interaction networks retrieved from IntAct and AURA DB, respectively. This approach provided us with thousands of possible 3'UTR-protein complexes, and as expected, we observed that **EMFs are more likely to form complexes than other proteins** (Fisher's exact test pval = 8.54e-46, odds ratio = 8.84), with 192 EMFs out of 238 forming at least one complex.

To filter out less likely candidate complexes, we further selected complexes which pass the following conditions: a) all components (3'UTR, nascent protein, RBP and intermediate protein) being present in at least one same tissue using the Human Protein Atlas dataset; b) the RBP does not interact with all the alternative 3'UTR forms of a given target protein, thus possibly regulating the nascent protein dependent on the 3'UTR present; c) the nascent protein is present in at least two dissimilar cellular components, using PrOnto (Chapple, Herrmann and Brun, 2015) probabilities for dissimilar GO terms. With this stringent filtering **we obtain**

191 distinct complexes comprising 42 RBPs and 116 intermediate proteins thereby predicted to affect the subcellular localisation and possibly the multifunctionality of 27 out of 238 EMF proteins (as nascent proteins in the complex).

Conclusion and perspectives

Overall, this analysis allows us to estimate and decipher the prevalence of an ill-known regulation mechanism and evaluate its role on protein multifunctionality. Importantly, multifunctionality has to be fully understood because (i) it is implicated in the regulation of the biological processes through their coordination and switch, (ii) it contributes to the complexity of the genotype-phenotype relationship by blurring and diversifying phenotypes, and (iii) it causes drug side-effects due to interferences with drug-target undisclosed function.

We will further determine the likelihood and experimentally validate several of 3'UTR-protein complexes, as well as analyse extensively the biological functions that may be regulated by this mechanism.

Finally, motivated by the results of our first model, we will expand our search of 3'UTR-dependent protein complexes to the whole human interactome, with the aim to ascertain the extent of the presence of these complexes in the cell in general, outside the context of multifunctionality.

References:

Berkovits, B. D. and Mayr, C. (2015) "Alternative 3' UTRs act as scaffolds to regulate membrane protein localization.," Nature, 522, pp. 363–367. doi: 10.1038/nature14321.

Chapple, C. E. et al. (2015) "Extreme multifunctional proteins identified from a human protein interaction network.," Nature communications, 6, p. 7412. doi: 10.1038/ncomms8412.

Chapple, C. E., Herrmann, C. and Brun, C. (2015) "PrOnto database: GO term functional dissimilarity inferred from biological data.," Frontiers in genetics, 6, p. 200. doi: 10.3389/fgene.2015.00200.

Mayr, C. (2016) "Evolution and Biological Roles of Alternative 3'UTRs," Trends in Cell Biology. Elsevier Ltd, 26, pp. 227–237. doi: 10.1016/j.tcb.2015.10.012.

Mayr, C. (2017) "Regulation by 3'-Untranslated Regions.," Annual review of genetics, 51, pp. 171–194. doi: 10.1146/annurev-genet-120116-024704.

Zanzoni, A., Chapple, C. E. and Brun, C. (2015) "Relationships between predicted moonlighting proteins, human diseases, and comorbidities from a network perspective.," Frontiers in physiology, 6, p. 171. doi: 10.3389/fphys.2015.00171.

Keywords: protein multifunctionality, protein moonlighting, 3'UTR regulation, interaction networks

Enhancer-gene associations in complete genomes unravel ancestral vertebrate regulation and key principles of enhancer function

Yves Clement * [1], Patrick Torbey [1], Pascale Gilardi-Hebenstreit [1], Hugues Roest Crollius [1]

[1] Institut de biologie de lÉNS Paris (UMR 8197/1024) (IBENS) – École normale supérieure - Paris, Institut National de la Santé et de la Recherche Médicale : U1024, Centre National de la Recherche Scientifique : UMR8197 – 46, Rue d'Ulm75005 Paris, France

Enhancers are short DNA sequences that bind transcription factors and contact promoters in *cis* to activate or repress the transcription of genes into RNA [1]. This control - or regulation - of gene expression by enhancers ensures the fine tuning of protein abundance in cells. The rise of next generation sequencing technologies has enabled large-scale epigenomics projects to map regulatory regions in a genome (e.g. ChIP-seq, ATAC-seq, DNAse1) [2], but these studies remain at the descriptive level and do not enable us to study the mechanisms of gene regulation because assigning an enhancer to their target gene(s) remains a difficult task.
Using a simple "nearest gene" approach will lead to incorrect assignments as enhancers can regulate genes over long distances (typically several hundreds of kilobases). Methods have been developed to map long distance regulatory interactions genome-wide, e.g. based on chromosomal conformation capture [3,4]. These experimental methods are complex to set up and identify regulatory interactions that are specific to a particular cell type or tissue, i.e. hardly transferable to other biological contexts.

We previously introduced PEGASUS (Predicting Enhancer Gene Associations Using Synteny), a computational method to assign target genes to enhancers, using computation predictions based on synteny, restricted to the human X chromosome [5]. This method works in a cell-type of tissue agnostic manner and relies on the analysis of evolutionary signals rather than on a costly and labour-intensive experimental setup. Here, we applied this method on the entire human and zebrafish genomes (no such regulatory map exists in the latter). We associated _˜1,300,000 predicted enhancers to _˜18,000 target genes in human and _˜55,000 predicted enhancers to _˜17,000 target genes in zebrafish. By comparing human and zebrafish predictions, we outlined a set of _˜600 genes in human and zebrafish with conserved *cis*-regulation in vertebrates, which are enriched in brain and development functions. We found in the human genome evidence for a direct link between the number of genes associated with a gene and the number of tissues this gene is expressed in. Finally, we found that the average distance separating enhancers and their target genes scales with genome size, showing that little selective pressure acts to preserve this distance.

Our collection of predicted enhancer-gene associations will facilitate and improve genetic and

*Speaker

genomic studies and lead to more precise mechanistic hypotheses, for example linking regions identified as active in a particular cellular context to their target genes or annotating genetic variants associated with a disease.

1. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014;15: 272–286. doi:10.1038/nrg3682

2. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

3. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015;47: 598–606. doi:10.1038/ng.3286

4. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167: 1369–1384.e19. doi:10.1016/j.cell.2016.09.037

5. Naville M, Ishibashi M, Ferg M, Bengani H, Rinkwitz S, Krecsmarik M, et al. Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome. Nat Commun. 2015;6: 6904. doi:10.1038/ncomms7904

Keywords: enhancer, gene expression, human, zebrafish, comparative genomics

Multiple probabilistic models resolve the functional organization of the cryptochrome/photolyase protein family

Riccardo Vicedomini *† [1,2], Jean-Pierre Bouly [1], Angela Falciatore [1], Alessandra Carbone‡ [1,3]

[1] Laboratoire de Biologie Computationnelle et Quantitative (LCQB) – Sorbonne Université UPMC Paris VI, Centre National de la Recherche Scientifique - CNRS – France

[2] Institut des Sciences du Calcul et des Données (ISCD) – Sorbonne Université UPMC Paris VI – France

[3] Institut Universitaire de France (IUF) – Ministère de l'Enseignement Supérieur et de la Recherche Scientifique – France

Context

Recent progresses in genomics and metagenomics are providing unprecedented opportunities to discover novel light-dependent proteins. New variants, exhibiting different spectral tuning or diversified functions, have been discovered in aquatic organisms, changing current views on their evolution (Jaubert *et al.*, 2017). An impressive diversification was in particular found within the cryptochrome/photolyase family (CPF), composed of flavoproteins displaying similar structures but a large variety of functions (Chaves *et al.*, 2011; Fortunato *et al.,* 2015; Sancar 2003): photolyases, blue-light activated enzymes that repair UV-induced DNA damages (CPD and (6-4) lesions), and cryptochromes, known for their multiple functional roles such as photoperception, photomagnetoreception, light-induced stress response. Cryptochromes with light-independent activity have also been found as part of the central circadian oscillator (Fortunato *et al.,* 2015). However, some CPF proteins have still unclear characterized function (e.g., the widespread Cry-DASH sub-family).

Although a lot of progress has been made, our ability to uncover the mechanisms underlying functional diversity of the CPF is still limited. Computational approaches exploiting the fact CPF proteins share a common domain organization and structural properties have been tried with no success. In most cases, the function of newly identified proteins cannot be anticipated with tools usually employed for phylogenetic reconstruction. Moreover, the original distinct separation of gene-regulating cryptochromes and DNA-repairing photolyases is gradually vanishing, as there are now several examples of CPF members exhibiting both functions (Coesel *et al.*, 2009; Heijde *et al.*, 2010). For these reasons, CPF represents a great challenge to test the method we developed for functional characterization.

Methods

*Speaker

†Corresponding author: riccardo.vicedomini@upmc.fr

‡Corresponding author: alessandra.carbone@lip6.fr

A reasonable way to organize sequences is through their predicted domains. Widely used domain search methods (Altschul *et al.*, 1997; Eddy, 2011) are based on a mono-source annotation strategy, where a single probabilistic model, generated from the consensus of a set of homologous sequences, is used to represent a protein domain. The mono-source strategy usually performs well as consensus models capture the most conserved features in domain sequences. However, when sequences have highly diverged, consensus signals become too weak to generate a useful probabilistic representation and global-consensus models do not characterize domain features properly. By generating multiple probabilistic models for a domain, describing the spread of evolutionary patterns in different phylogenetic clades, we can effectively explore domains that are likely to be coded in gene sequences spanning the entire phylogenetic tree of species and possibly presenting remote homology. More precisely, we exploit a recent advance in domain annotation (Bernardes *et al.*, 2016; Ugarte *et al.*, 2018) to design a new computational strategy and enrich the CPF sequences by finding new members and possibly predicting their functions.

Recently, the genome-based domain annotation tool CLADE introduced a *multi-source* strategy (Bernardes *et al.*, 2016) in which protein domains are represented by a large number of probabilistic models. More in detail, it considers all sequences associated to a domain family of Pfam (Finn *et al.*, 2014) and, for some representative query sequences, it constructs a *clade-centered* model (CCM) by retrieving a set of homologous sequences close to each query sequence. The main idea is that CCMs display features that are characteristic of the query sequence. Therefore, the more domain sequences are divergent, the more CCMs are expected to highlight different features. In our work, we use the power of the multi-source annotation strategy to look at CPF sequences from the point of view of multiple models in the hope that their profiles could highlight functional characteristics of the sequence. However, the idea we present was not specifically tailored to the CPF and its FAD-binding domain region. Therefore, it could be applied in general to any protein (or domain) family.

The approach we propose consists on three main phases: the construction of a CLADE-like CCM library, the mapping of the library and the contextual construction of a multidimensional space by assigning a *feature vector* to each sequence we want to characterize, and a hierarchical clustering strategy.

As opposed to CLADE, we build a CCM library in a different manner. First, as specific signatures related to the functions are usually found in the FAD-binding region of CPF proteins, we focus exclusively on a *single* Pfam domain (*i.e.*, the FAD-binding domain of DNA photolyase). In this way, we can also afford to build a probabilistic model from each domain sequence belonging to Pfam in reasonable time. Second, we build the CCM library as a collection of profile Hidden Markov Models (pHMMs), instead of position-specific scoring matrices, in order to use a more powerful formalism. This task is carried out with hhblits (Remmert *et al.*, 2012) for searching homologous protein sequences and HMMER (Eddy, 1998) for mapping pHMMs to protein sequences. More precisely, CCMs are built looking for UniProt sequences having at least 60% of identity with respect to the query in order to possibly preserve those motifs that might be related to a specific function.

Each CCM is finally mapped to the sequence in order to look at it from the point of view of the models. In particular, we use the *match score* reported by HMMER for each pair model-sequence. This is done for all the sequences we want to classify. Then we discard sequences that have partial matches against most of the models (and that might be incomplete) and CCMs based on very few sequences. We define, for each CPF sequence S that we retained, a feature vector where each component consists of the score of a model against S. In this way, each sequence is described as a point in a m-dimensional space, where m is the size of the model library. These points are then clustered using a hierarchical agglomerative strategy that allows us to define what we call *function tree* (FT).

Finally, we defined a way to select the model that better describes a specific sub-tree/cluster in FT. This is done by considering the CCM achieving the highest scores on most of the sequences of the cluster and leaving all the others with lower scores. This model will help us to identify conserved positions specific to a sub-tree, if any.

Results

We applied the aforementioned computational strategy to a set of 397 CPF protein sequences which were selected in order to span as much as possible the whole tree of life. Moreover, 69 of them were already functionally characterized.

A classical distance tree based on the full-length sequence finds three main groups exhibiting disparate functions within them: (i) plant photoreceptors (pCry), pCry-Like and CPD photolyases; (ii) the Cry-DASH; (iii) (6-4) photolyases, insect photoreceptors, and light independent cryptochromes.

In contrast, our function tree resolves major groups with a coherent functional organization.

Namely, first, it separates light-independent transcription regulator cryptochromes from the light-dependent (6-4) photolyases and photoreceptors in superclass (iii). Second, it separates CPD photolyases from photoreceptors in superclass (i).

These encouraging results show that our computational approach might be effectively used for anticipating functions but also for revealing the existence of novel light-sensitive proteins. In this respect, we clearly identified a subset of proteins (we named Ncry) that are very close to the CPD photolyases in the distance tree but that, with respect to the function tree, are completely separated and put much closer to other sequences with different known function. In order to perform a preliminary validation of the function tree, we aligned and compared the two models that better describe CPD photolyase and Ncry sequences in the function tree. As a result, we assessed that most of conserved amino acids already known to be *necessary* for the CPD functional activity on the former were completely missing in the latter. We were also able to highlight some conserved positions and motifs between model profiles characterizing sequences far apart in the distance tree yet exhibiting the same function according to the literature. Hence, an exhaustive comparison of all models that better characterize each identified group of the function tree is likely to reveal conserved residues (or physico-chemical properties) that may play a crucial role with respect to the protein function. Therefore, our method could not only partition a set of sequences in a functionally coherent manner but also it could be extremely useful for extracting functionally important motifs from the sequences.

References

Jaubert *et al.* Light sensing and responses in marine microalgae. Current Opinion in Plant Biology. 2017;37:70-77.
Chaves *et al.* The cryptochromes: blue light photoreceptors in plants and animals. Annual review of plant biology. 2011;62:335-364.

Fortunato *et al.* Dealing with light: the widespread and multitasking cryptochrome/photolyase family in photosynthetic organisms. Journal of plant physiology. 2015;172:42-54.

Sancar A. Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors. Chemical reviews. 2003;103(6):2203-2238.

Coesel *et al.* Diatom PtCPF1 is a new cryptochrome/photolyase family member with DNA repair and transcription regulation activity. EMBO reports. 2009;10(6):655-661.

Heijde *et al.* Characterization of two members of the cryptochrome/photolyase family from Ostreococcus tauri provides insights into the origin and evolution of cryptochromes. Plant, cell environment. 2010;33(10):1614-1626.

Altschul *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.

Eddy SR. Profile hidden Markov models. Bioinformatics (Oxford, England). 1998;14(9):755-763.

Bernardes *et al.* Improvement in protein domain identification is reached by breaking consensus, with the agreement of many profiles and domain co-occurrence. PLoS computational biology. 2016;12(7):e1005038.

Ugarte *et al.* A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. 2018. Submitted.

Finn *et al.* Pfam: the protein families database. Nucleic Acids Research, Volume 42, Issue D1, 1 January 2014, Pages D222–D230.
Michael Remmert *et al.* HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods volume 9, pages 173–175 (2012).

Keywords: cryptochrome/photolyase family, protein function characterization, probabilistic models, clade, centered models, conserved motifs extraction

Statistical modeling of bacterial promoter sequences for regulatory motif discovery using expression data

Ibrahim Sultan *† 1, Sophie Schbath 1, Pierre Nicolas 1

1 INRA-MaIAGE – Institut national de la recherche agronomique (INRA) : UR1404 – France

Transcription factors play a key role in mediating the adaptation of bacteria to environmental conditions. Powerful algorithms and approaches have been developed for the discovery of their binding sites but automatic de novo identification of the main regulons of a bacterium from genome and transcriptome data remains a challenge. The approach that we propose here to address this task is based on a probabilistic model of the DNA sequence that can make use of precise information on the position of the transcription start sites and of condition-dependent transcription profiles. Two main novelties of our model are to allow overlaps between motif occurrences and to incorporate covariates summarizing transcription profiles into the probability of occurrence in a given promoter region. Each covariate may correspond to the coordinate of the gene on an axis (e.g. obtained by PCA or ICA) or to its position in a tree (e.g. obtained by hierarchical clustering). All the parameters are estimated in a Bayesian framework using a dedicated trans-dimensional MCMC algorithm. This allows simultaneously adjusting, for many motifs and with many transcription covariates, the width of the corresponding position weight matrices, the number of parameters to describe positions with respect to the transcription start site, and the covariates that are relevant. Results obtained for the bacterium Listeria monocytogenes are presented.

The mathematical equations describing the model, and a figure demonstrating the results, are included in the uploaded file.

Keywords: Bioinformatics, Mixture models, Stochastic algorithms, Bayesian methods.

[*] Speaker

[†] Corresponding author: islam.sultan@inra.fr

Mathematical modeling of Fe-S biogenesis shows strong links between iron homeostasis and oxidative stress response

Firas Hammami * [1,2], Frédéric Barras [3], Pierre Mandin† [1], Elisabeth Remy‡ [2]

[1] Laboratoire de chimie bactérienne (LCB) – Aix Marseille Université : UMR7283, Centre National de la Recherche Scientifique : UMR7283 – 31 Chemin Joseph Aiguier 13402 MARSEILLE CEDEX 20, France

[2] Institut de Mathématiques de Marseille (I2M) – Aix Marseille Université : UMR7373, Ecole Centrale de Marseille : UMR7373, Centre National de la Recherche Scientifique : UMR7373 – Centre de Mathématiques et Informatique (CMI)Technopôle Château-Gombert39, rue Frédéric Joliot Curie13453 Marseille Cedex 13, France

[3] Institut Pasteur – CNRS : ERL6002 – 25-28 Rue du Dr Roux, 75015 Paris, France

Iron-sulfur (Fe-S) clusters are essential cofactors conserved in all domains of life, but extremely sensitive to stresses such as iron deprivation and oxidative stress. Thus, to control those stresses, adaptation mechanisms relying on complex regulatory networks prevent from reactive oxygen species production and adjust free intracellular iron levels..

To get a better understanding on how environmental conditions modulate Fe-S biogenesis in the bacterium *E. coli*, we used a logical mathematical model. This modeling approach consists in a directed signed graph representing the regulations (activations or inhibitions) between components, and logical rules attached to each node depicting its dynamical behaviour with respect to the state of its regulators. We constructed a logical model centered on three modules describing the molecular actors acting on Fe-S cluster biogenesis : the Fe-S biogenesis module containing the Fe-S cluster assembly machineries Isc and Suf and the IscR transcription factor (TF), the main regulator of Fe-S homeostasis ; the iron homeostasis module containing the free intracellular iron regulated by the iron sensing TF Fur, repressing iron import genes, and the non-coding regulatory RNA RyhB involved in iron sparing ; and the oxidative stress module composed of the Reactive Oxygen Species, able to activate the H2O2 sensing TF OxyR, and enabling catalase and iron sequestrating proteins expression in order to decompose H2O2 and limit Fenton reaction[2].. Inputs of the model represent extracellular iron and oxygen environments conditions, and ErpA and NfuA proteins, able to carry the cluster to Fe-S proteins, are outputs nodes of the model.

The modular structure of the regulatory graph emphasizes the interactions between Fe-S biogenesis, iron homeostasis, and oxidative stress response (ROS) modules. While the model shows one attractor per condition, the number of oscillating modules increases with iron and oxygen;

*Speaker

†Corresponding author: pmandin@imm.cnrs.fr

‡Corresponding author: elisabeth.remy@univ-amu.fr

as a consequence, the system attractor size increases also. This observation may reflect a greater need of adaptation of the bacteria when both iron and oxygen accumulate. Moreover, the Fe-S biogenesis module reveals two main regimes: an homeostatic regime generated by the Isc and Suf machineries, and a stress regime, where only the Suf machinery is active. The model helps in classifying four different behaviors of Suf expression, depending on environmental conditions.

Altogether mathematical modeling gives us a framework where we can predict Fe-S biogenesis genes behavior regardless of iron and oxygen levels. Moreover, the deep links between ROS and iron homeostasis modules suggest that the combination of the two signals control Fe-S biogenesis.

(Work in progress)

References

[1] Roche B, Aussel L, Ezraty B, Mandin P, Py B, Barras F. Iron/sulfur proteins biogenesis in prokaryotes: formation, regulation and diversity. *Biochim Biophys Acta,* Mar;1827(3):455-69, 2013

[2] Imlay, J.A.: The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium. Nature Reviews Microbiology 11(7), 443–454 (2013).

Keywords: Bacteria, mathematical modelling, Fe, S cluster biogenesis, iron homeostasis, ROS, oxidative stress response

Biocuration and rule-based modelling of protein interaction networks in KAMI

Sébastien Légaré * , Ievgeniia Oshurko [1], Russel Harmer† [1]

[1] Laboratoire de l'Informatique du Parallélisme (LIP) – École Normale Supérieure - Lyon – France

Summary:

KAMI, the Knowledge Aggregator and Model Instantiator, is a software for biocuration and modelling of molecular interaction networks. It provides a knowledge representation to unambiguously express the details of biomolecular interactions. This representation can be built either programmatically or graphically via the KamiStudio interface. To assist users in curating their biological knowledge, KAMI is organised in two distinct layers: a network and a set of individual interactions called nuggets. Once a new nugget is built, it can be automatically aggregated to the network. The software then performs a series of tests to ensure consistency including duplicate search, biological database grounding and semantic checking. This greatly facilitates biocuration as users do not need to have the complete network in mind to add new data. Furthermore, interaction networks represented in KAMI can be directly converted to rule-based models in the Kappa language for simulation and analysis. In this talk, we will present the use of KAMI through a model of tyrosine phosphorylation involved in cell signaling. This example is well suited to showcase the advantages of the rule-based strategy. In particular, we will demonstrate the use of causality analysis to discover pathways in the model that were not explicitly input by the user.

Knowledge representation:

At the heart of KAMI [1] is its knowledge representation. Knowledge about biomolecular interactions is represented in the form of graphs. Nodes represent protein components and interactions, while edges are relationships between them. Protein component nodes can be either a protein itself or a structural element of the type region, site or residue. Interactions can be either modifications or bindings. Modifications encompass any change in the state of a protein component, from phosphorylation to conformational change. Two additional nodes types, the test and the state, allow the representation of conditions for interactions to occur. Users can hence express complex interactions with a high level of structural and mechanistic detail. For example, the regions through which two proteins bind can be specified. The chosen set of node types along with their edge connectivity readily translates to rules in the Kappa language [2], as will be discussed further below.

The knowledge representation can be built either programmatically or graphically. Users can

*Speaker

†Corresponding author: russell.harmer@ens-lyon.fr

choose to write/read their interactions in Python language using the KAMI library or to draw/visualize them graphically from the KamiStudio interface. The complete equivalence between those two input methods is made possible by the internal data structure of KAMI, which is itself a graph built with NetworkX [3]. Typically, the programmatic input would be used to add interactions in batch mode and the graphical interface to enter detailed interactions one by one.

A pivotal aspect of KAMI is that it splits biological knowledge in two distinct layers: a network and a set of individual interactions called nuggets. Once an interaction nugget is created, either programmatically or graphically, it can be automatically aggregated to an existing network. However, the nugget is not simply blended in the network. It is instead kept as an individual interaction and the matching between its nodes and the equivalent nodes in the network is stored as a graph homomorphism. Users can then keep track of each single interaction they added to the network and subsequently modify or remove them individually if needed. This greatly facilitates biocuration by allowing the incremental aggregation of knowledge as will be elaborated below.

Other graphical knowledge representations for molecular interactions exist, the most notable probably being SBGN [4]. In essence, KAMI's representation resembles the SBGN entity relationship diagram. But important differences subsist at the practical level. SBGN relies on a myriad of different symbols and the combination of three types of diagrams to fully disambiguate molecular interactions. KAMI on the opposite uses a relatively limited amount of symbols and directly represents interactions unambiguously. It does so by allowing an interactive navigation between its two layers of knowledge, the network and nuggets described above. In our opinion, that makes KAMI much superior for use on digital media, although less amenable to paper media.
KAMI can be seen as a rule-based analog of CellDesigner [5]. In the latter, the interaction network resembles a Petri net where nodes are molecular species and edges are transitions between them. With our approach, the network rather is a contact map where nodes are individual molecules and edges are their interactions, with the underlying nuggets resolving potential ambiguities. However, KAMI is not simply a tool to build Kappa models graphically. It also aims at providing a framework for biocuration, the continuous development of biomolecular networks.

Biocuration:

The construction of molecular interaction networks or biological models in general comes with several difficulties. As a model grows, it becomes increasingly tedious to add new data in a consistent manner. Every addition or refinement must be checked to ensure that it does not contradict, duplicate or incorrectly interact with existing parts of the model. To solve that issue, KAMI features an "aggregation engine" that performs checks each time a new nugget is aggregated to a network. It searches for protein components that are already present in the network. Once components in common between the new nugget and the network are found, the new interaction can be matched in the network if it already exists. This way, users can see if the data they add is new with respect to a given network or if it duplicates, modifies or adds details to an existing interaction.

The aggregation engine is optimized for biologically grounded data. It is hence advantageous to refer to protein components using standard identifiers from databases like HGNC, UniProt, Ensembl, InterPro, etc. To help find biological grounding for nuggets, KAMI includes an "anatomizer" to fetch the "anatomy" of proteins from online databases. This includes the various standard identifiers but also regions, sites, post-translational modifications and more. Proper biological grounding is however not mandatory as users may want to introduce hypothetical interactions taken from their own experiments.

KAMI also performs semantic checking during nugget aggregation. This serves to avoid accidentally introducing interactions that are against common scientific knowledge. For example, a warning would be ensued if a user created a nugget where a tyrosine kinase phosphorylated a threonine. This is implemented through what is referred to as ”semantic nuggets”. This type of nugget takes the same form as the interaction nuggets described previously, but their nodes refer to generic components rather than specific proteins. Creating a semantic nugget is akin to making the software learn about the general properties of biological systems. While aggregating an interaction nugget, KAMI tries to match it with its available semantic nuggets. If it finds a match, the software in a sense ”understands” what the interaction means. It can then advise the user if the nugget contradicts general principles of biology. Semantic nuggets are viewed as elements that will be added over time by the developers of KAMI rather than by users.

Rule-based modelling:

As discussed above, KAMI's knowledge representation can be converted to an executable model in the Kappa language [2]. The conversion is straightforward thanks to the rule-based strategy. Effectively, every nugget corresponds to a single Kappa rule or set of rules that is independent of other nuggets. If a reaction-based strategy had been adopted instead, each nugget would potentially combine with every other nugget that share some protein to produce a combinatorially large number of reactions. In the rule-based setting, species arising from combinations of rules occur naturally in simulation with a propensity that reflects their probability of occurrence.

Rule-based modelling also allows the specification of resources, or sites, through which interactions occur. This means that the structural information entered at the level of KAMI is taken into account in the model. If two molecules are known to bind to the same site of a target protein, it can be represented in the network by pointing the edges from both molecules to the same site on the protein. In that case, competition between the two molecules would automatically arise during the rule-based simulation.

KAMI is intended to allow a quantitative study of biomolecular networks. A rule-based setting was hence chosen rather that a Boolean or multi-valued modelling language like GINsim [6]. Kappa, the chosen rule-based language, shares a lot of similarity with BioNetGen [7]. Indeed, nothing prevents us from adding it as a future output. However, Kappa offers analysis tools like causality analysis which are, to our knowledge, unavailable with other rule-based modelling languages.

Analysis:

Once a KAMI interaction network is converted to a Kappa rule-based model, it can be simulated with KaSim. Those simulations provide the amount of each species of the model as they evolve over time according to their interactions. They can be analysed like any other type of dynamic simulation to study dose-response relationships, oscillatory behaviors, multistability, etc. Additionally, analyses specific to rule-based modelling can be performed using tools from the Kappa software suite. These include analysis of polymerization, dynamic influence maps and causality. In particular, the causality analysis allows users to find pathways in the network. Recall that in rule-based modelling, users just need to input fundamental rules and the combination of these rules occurs spontaneously in simulation. This means that causal traces can be used to actually discover pathways in the model that the user had not foreseen. This is a radically different way to use models compared to what can be done with a reaction based approach, where every pathway must be explicitly written and simulations often serve just to confirm hypotheses.

The tyrosine signaling model:

To demonstrate the use of KAMI and rule-based modelling, we will present a model of tyrosine phosphorylation involved in human cell signaling. This model includes tyrosine kinases, their targets, and proteins containing SH2 domains which bind to tyrosines once they were phosphorylated. Interactions between those three types of molecules were gathered from databases PhosphoSite [8], Phospho.ELM [9], the NCI PID [10] and the literature [11-12]. The interactions were converted to KAMI nuggets and aggregated to form a network of 1185 interactions across 175 different proteins. This model is representative of general signaling networks and is well suited to showcase the advantages of rule-based modelling. In particular, some of the proteins contain both a kinase and a SH2 domain. These proteins can bind to other proteins or complexes through their SH2 domain and then quickly phosphorylate multiple tyrosines within the complex. The formation of large protein complexes through the combination of binding rules then becomes crucial if one wants to reproduce the dynamic behavior of the system.

Future developments:

Coming improvements to KAMI will include a versioning system that will allow users to see the history of their interaction networks and revert any changes. A database version of the knowledge representation will also be developed so that members of a team can work efficiently on a same project from different locations. Node transitivity will be used to automatically recognize when a new nugget is actually a more detailed version or simplification of an existing interaction. A tool will be added to manage protein splice variants. Finally, we plan to add the possibility to build interactions using intuitive sentences in addition to the already existing programmatic and graphical inputs.

Perspectives:

We believe that KAMI will significantly ease the biocuration of complex molecular interaction networks and leverage the power of rule-based modelling. This will allow systems biologists to truly model systems, rather than being restricted to modelling specific pathways.

References:

1) Harmer R., Le Cornec Y.-S., Légaré S. and Oshurko I. CMSB, 2017, LNBI 10545:3-19
2) kappalanguage.org
3) networkx.github.io
4) Le Novère et al., Nature Biotechnology, 2009, 27:735, doi:10.1038/nbt.1558
5) www.celldesigner.org
6) ginsim.org
7) Harris et al. Bioinformatics, 2016, 32:3366-3368
8) www.phosphosite.org
9) phospho.elm.eu.org

10) github.com/NCIP/pathway-interaction-database/tree/master/download
11) Schulze W. X., Deng L. and Mann M. Molecular Systems Biology, 2005, 25 May, doi:10.1038/msb4100012
12) Liu et al. Cell Communication and Signaling, 2012, 10:27, doi:10.1186/1478-811X-10-27

Keywords: Biocuration, knowledge representation, rule, based modelling, protein interaction network, cell signaling, phosphotyrosine

Build your own multi-omics website with BACNET

Christophe Becavin * [1], Nicolas Maillet [1], Pierre Lechat [1]

[1] Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 25-28 Rue du Docteur Roux, 75015 Paris, France

To face up to the exponential growth of heterogeneous omics datasets of various organisms and the lack of personalised user interface dedicated to biologists we developed BACNET, a Java based platform for rapid development of multi-omics website and desktop application with genome viewer, heatmap viewer and many other tools.
There are already many referent databases for each type of "omics" that exist. The development of these databases was mandatory to allow a reproducible science. However most of them are only repositories for raw datasets and metadata information. Extensive bioinformatic is needed to make use of the data. Consequently, most of the biologists will not be able to access or analyse them because of lack of time or competence. There is a real need for rich user web-interface that would allow biologists to go through these datasets and have an in-situ vision of the studied processes.

For big multi-omics project like ENCODE and TCGA, and also for well-known model organisms like the Yeast and E.coli such websites exist. In those, one can browse the datasets in a specialised genome browser, or can perform a search by gene expression, biological condition or protein presence. But with the decrease of the sequencing price, there are more and more multi-omics studies produced on specific organisms. We observed this phenomenon at the Institut Pasteur, where each bacterium and parasite is now being studied with multi-omics approaches. Biologists in the concern laboratories need these user interfaces to make sense of their datasets. Not only desktop software like the Integrated Genome Browser but also website to share their results in the organism related category.

We started in 2012 the development of a fully functional "omics" data analysis platform named BACNET with a heatmap tool, and a genome viewer. This platform is based on Eclipse RCP and allowed us to construct a desktop application with an interactive Graphic User Interface (genome zoom features, data selection, search element, etc.). We used Eclipse RCP and RAP for the Graphic User Interface (GUI). Eclipse RCP (Rich Client Platform) is a powerful Java tool for building easily rich GUI for application. The difference between both is that RCP helps to develop software on MAC or PC, whereas RAP is for developing powerful website. Consequently, thanks to the same Java code, our platform can work on every MAC or PC for local analysis of the data, and on a website to be used for data publication.

The first GUI we developed was a Heatmap management tool. It allows displaying a table of expression, colouring its columns according to the value of expression, filtering and ordering

*Speaker

rows or columns, exporting to image and text file, and common statistical tools analysis. This Heatmap management tool has been used by several people in Pascale Cossart's laboratory at Institut Pasteur for Listeria transcriptomic analysis but also for other type of "omics" studies. The second GUI we developed is a genome viewer displaying genome information and all transcriptomic data we already described. Genome tracks can be browsed, zoomed, overlaid, or displayed separately. Thanks to this architecture we now have a genome viewer in which every gene expression array, tiling array, RNASeq data, and proteomics data can be displayed and compared. We further developed the platform by adding new capabilities thanks to the development of different multi-omics project in our laboratory. We added for example tools for proteomics datasets management and non-coding RNA. In the summer 2014, we started to use BACNET platform to develop a website specifically design for listeriologists with highly dynamic user-interface for analysing and browsing every 'omics' data available for the Listeria species. Most of the efforts were put into the curation of all datasets include in this website, and in providing a user-friendly interface.

Listeria monocytogenes is a foodborne pathogen responsible for foodborne infections with a mortality rate of 25%. Over the past three decades Listeria has become a model organism for host-pathogen interactions, leading to critical discoveries in a broad range of fields, including virulence-factor regulation, cell biology, and bacterial pathophysiology. To study these mechanisms, several genomics, transcriptomics, and proteomics data sets have been produced.

We have developed a web-based platform, named Listeriomics (http://listeriomics.pasteur.fr/), that integrates the different BACNET tools (see Figure) for omics data analyses, i.e., (i) an interactive genome viewer to display gene expression arrays, tiling arrays, and sequencing data sets along with proteomics and genomics data sets; (ii) an expression and protein atlas that connects every gene, small RNA, antisense RNA, or protein with the most relevant omics data; (iii) a specific tool for exploring protein conservation through the Listeria phylogenomic tree; and (iv) a coexpression network tool for the discovery of potential new regulations.

To our knowledge, none of the referent databases dedicated to model organisms such as *E. coli* or *B. subtilis* integrates as many data sets and visualization tools as the Listeriomics resource does. User experience and feedback from our collaborators using the Listeriomics interface for the past 5 years were driving forces in organizing and improving the way to access data and tools. Our main purpose was to design an easy-to-use website with a dynamic interface for biologists wanting to access the different heterogeneous "omics" data sets available for Listeria. As such our website should interest the JOBIM community as it shows an example of extensive multi-omics data integration for model organism.

In the last two years, BACNET platform has been used in the bioinformatic HUB to create different websites related to multi-omics analysis performed within the Institut Pasteur. Two of the websites created will soon be published. One is related to a multi-omics analysed of the parasites Leishmania, the other is dedicated to the bacteria Yersinia. Two pathogenic organisms with critical global health implication but with no centralised database or website for visualizing their different omics scale. Thanks to BACNET we easily build these websites using standard sequencing and tab-delimited files.
The BACNET platform is still in development (https://gitlab.pasteur.fr/bacnet). We are improving the code to assure that every bioinformatician wishing to create their own multi-omics website for another organism can do so with few efforts in one day. We believed the description of the BACNET platform will interest the JOBIM community.

Keywords: Personalized website, genomics, transcriptomics, proteomics, systems biology

Gigwa - Genotype Investigator for Genome-Wide Analyses

Guilhem Sempéré* [1,2], Adrien Petel [1], Alexis Dereeper [2,3], Manuel Ruiz [2,4], Pierre Larmande †‡ [2,5]

[1] UMR Intertryp - CIRAD - IRD – Institut de recherche pour le développement [IRD], CIRAD – Avenue Agropolis - 34398 Montpellier Cedex 5, France
[2] Institut de Biologie Computationnelle (IBC) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut National de la Recherche Agronomique, Institut National de Recherche en Informatique et en Automatique, Université de Montpellier, Centre National de la Recherche Scientifique – Building 5 - 860 rue de St Priest 34095 Montpellier, France
[3] IRD IPME (IPME) – Institut de recherche pour le développement [IRD] – Avenue Agropolis, 34398 Montpellier Cedex 5, France
[4] Centre de coopération internationale en recherche agronomique pour le développement (CIRAD) – CIRAD, Institut National de la Recherche Agronomique - INRA – Av Agropolis, Montpellier, France
[5] Institut de Recherche pour le Développement (IRD) – Institut de recherche pour le développement [IRD] : UMR232, Université de Montpellier : UMR232 – 911 avenue Agropolis,BP 6450134394 Montpellier cedex 5, France

With the advent of next-generation sequencing (NGS) technology, thousands of new genomes of both plant and animal organisms have become available. In this context, the Variant Call Format (VCF) [1] has become a convenient and standard file format for storing variants identified by NGS / NGG approaches. VCF files may contain information on tens of millions of variants, for thousands of individuals. Having to manage such significant volumes of data involves considerations of efficiency with regard to the following aspects: Filtering features, Storage performance, Sharing capabilities, Graphical visualization. However, existing tools are often limited to command line or programmatic APIs targeted at experienced users, but are not suitable for non-bioinformaticians.

The Gigwa application [2], which stands for "Genotype Investigator for Genome-Wide Analyses", aims at taking into account those aspects. It provides an easy and intuitive way to explore large amounts of genotyping data by filtering it not only on the basis of variant features, including functional annotations, but also on genotype patterns. It is a fairly lightweight, web-based, platform-independent solution that allows to feed a MongoDB [3] NoSQL database with VCF [4], PLINK or HapMap files containing up to billions of genotypes, and provides a user-friendly interface to filter data in real time. Gigwa provides the means to export filtered data into several popular formats and features connectivity with visualization software such as FlapJack [5] and online or standalone genome browsers (GBrowse, [REF]JBrowse [6] and IGV [7]). Additionnally, Gigwa-hosted datasets are interoperable via two standard REST APIs: GA4GH[8] and BrAPI [9]. Thus, we think that Gigwa could serve a large number of scientists by helping them to manage, filter and share their own data.

*Corresponding author: guilhem.sempere@cirad.fr

†Speaker

‡Corresponding author: pierre.larmande@ird.fr

1. 1000 Genome project Consortium. Variant Call Format (VCF) [Internet]. [cited 2018 Mar 20].

2. Sempéré G, Philippe F, Dereeper A, Ruiz M, Sarah G, Larmande P. Gigwa-Genotype investigator for genome-wide analyses. Gigascience [Internet]. 2016 [cited 2016 Sep 24];5:25.

3. MongoDB Inc. MongoDB [Internet]. 2015 [cited 2015 Dec 19]. Available from: https://www.mongodb.org/

4. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics [Internet]. 2011 [cited 2014 Jul 10];27:2156–8.

5. Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas WTB, et al. Flapjack–graphical genotype visualization. Bioinformatics [Internet]. 2010 [cited 2016 Mar 3];26:3133–4.

6. Skinner ME, Uzilov A V, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. [Internet]. 2009;19:1630–8.

7. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. [Internet]. 2013;14:178–92.

8. The Global Alliance for Genomics and Health Consortium . GA4GH API [Internet]. 2017. Available from: https://github.com/ga4gh/ga4gh-schemas
9. Brapi consortium . The Breading API. 2017; Available from: https://brapi.org/

Keywords: Genomic variations, VCF, HapMap, NoSQL, MongoDB, SNP, INDEL

Molecular analysis and in silico research of expressed genes in relationship with salt stress tolerance in Medicago truncatula Gaertn.

Adel Amar Amouri * [1]

[1] Département de Biologie – Faculté des Sciences de la nature et de la vie. BP1524. Université d'Oran1 Ahmed Ben Bella. 31000 Oran., Algeria

In our study, we assessed the phenotypic variability of eleven ecotypes of *M. truncatula* Gaertn. under salt stress (137 mM NaCl) compared to the control at the germination stage. For the analysis of seedling growth under salinity stress, it will be useful to study root growth elongation. Several studies are focalized in root development because it is the most sensitive part of the plant and controls rapid transmission information to other plant parts. The Results showed that Tru 131 ecotype, with a high ratio (Root more vigorous than shoot) is more tolerant to salinity stress than the sensitive ecotypes. For the molecular analysis, four expressed sequence tag-simple sequence repeat EST-SSRs primers (MTIC 044, MTIC 124, MTIC 077 and MTIC 335) were used to show genetic variability in different ecotypes of *M. truncatula* Gaertn. comparing with the two contrasting genotypes Tru 131, tolerant genotype and Jemalong, sensitive one. The polymorphic information content (PIC) ranged from 0.12 to 0.49. These EST-SSRs markers were more polymorphic except MTIC 044. We have chosen the most polymorphic EST-SSR (MTIC124) in order to determine a potential link between this marker and salt stress tolerance. The results obtained from " Unigene and Uniprot" databases of highly similarity proteins sequences with the EST- SSR (MTIC 124), showed that this locus encode cysteine proteinase inhibitor, and was expressed principally in root in *M.truncatula.* This data suggest that this locus is involved in salinity tolerance, and it is appropriate for understanding salt stress tolerance mechanisms in *Medicago truncatula* Gaertn.

Keywords: Medicago truncatula Gaertn., Molecular databases (Unigene / Uniprot), Molecular markers (EST, SSR), Salt stress

*Speaker

A Novel Computational Approach for Global Alignment for Multiple Biological Networks

Warith-Eddine Djeddi *† [1], Sadok Ben Yahia *

[1], Engelbert Mephu Nguifo *

[2]

[1] LIPAH laboratory (LIPAH) – University of Tunis El Manar, 2092, Tunis, Tunisia, Tunisia
[2] Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS) – Université Clermont Auvergne, CNRS : UMR6158 – 1 rue de la chebarde, Campus universitaire des cézeaux, 63178 Aubière cedex, France

Due to the rapid progress of biological networks for modeling biological systems, a lot of biomolecular networks have been producing more and more protein-protein interaction (PPI) data. Analyzing protein-protein interaction (PPI) networks aims to find regions of topological and functional (dis)similarities between molecular networks of different species. The study of PPI networks has the potential to teach us as much about life process and diseases at the molecular level. The few methods that have been proposed in the for multiple PPI network alignment have some weaknesses. Thus, a new network alignment methods are of a compelling need. In this paper, we propose a novel algorithm for a global alignment of multiple protein-protein interaction (PPI) networks called MAPPIN. The latter relies on information available for the proteins in the networks, such as sequence, function and network topology. Our algorithm is perfectly designed to exploit current multi-core CPU architectures, and has been extensively tested on a real data (eight species). Our experimental results show that MAPPIN significantly outperforms NetCoffee in terms of coverage. Nevertheless, MAPPIN is handicapped by the time required to load the gene annotation file. An extensive comparison versus the pioneering PPI methods also show that MAPPIN is often efficient in terms of coverage, mean entropy or mean normalized entropy.
This work was recently accepted in the Journal IEEE/ACM Transactions on Computational Biology and Bioinformatics. The original paper could be found in the following link: http://ieeexplore.ieee.org/doc
Availability: Datasets and an implementation of the approach are freely available at https://www.isima.fr/meph

Keywords: Alignment, Biological networks, Computational biology, Gene ontology, Protein, protein interaction.

*Speaker

†Corresponding author: waritheddine@yahoo.fr

The genome of the Microsporidia Nosema granulosis, an endosymbiotic feminizing parasite of amphipod crustaceans.

Alexandre Cormier *† [1], Mohamed Amine Chebbi [1], Isabelle Giraud [1], Rémi Wattier [2], Maria Teixeira [2], Thierry Rigaud [2], Richard Cordaux [1]

[1] Laboratoire Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, Université de Poitiers, UMR CNRS 7267 – CNRS : UMR7267 – France

[2] Laboratoire Biogéosciences, Université Bourgogne Franche-Comté, UMR CNRS 6282 – CNRS : UMR6282 – France

Multicellular organisms have been continuously involved in complex interactions with microorganisms during their evolution, the most intimate of which is endosymbiosis. Over the past years, evidence has been accumulating that endosymbionts affect animal biology in many ways, such as host nutrition, development, immunity and even sex determination. In this last example, endosymbionts are able to disrupt the sex determination of hosts in favor of females because they are predominantly transmitted vertically through female egg cytoplasm. Current efforts in our laboratory are aimed to decipher genetic mechanisms underlying the ability of feminizing obligate intracellular endosymbionts (Microsporidia, fungi) to reverse genetic males into functional phenotypic females in a freshwater amphipod (*Gammarus roeselii*). We have generated high-throughput sequencing data (Illumina HiSeq2500) for *Nosema granulosis*, a Microsporidia species identified as vertically transmitted feminizing parasite infecting *G. roeselii*. Unlike other Microsporidia species, *N. granulosis* does not have any extracellular stage and cannot be isolated from host cells, thereby substantially complicating the sequencing and assembling of the genome. The sequencing data were assembled using SOAPdenovo2 after being cleaned with Trimmomatic. Identification of *Nosema* contigs was performed in two steps. First, genome assembly was filtered by identifying contigs sharing similarity sequence with 26 previously sequenced Microsporidia genomes and proteomes using Blast. Second, the sequences were taxonomically assigned using Blobtools. All contigs assigned to fungi were kept and form the *N. granulosis* genome. Genome completeness was assessed using BUSCO. After structural and functional annotation, we will compare the *N. granulosis* genome to the four available (non-feminizing) *Nosema* genomes: *N. antheraeae*, *N. apis*, *N. bombycis* and *N. ceranae*.

Keywords: Sex determination, Microsporidia, Genome assembly

[*] Speaker

[†] Corresponding author: alexandre.cormier@univ-poitiers.fr

Characterization of the drug resistances in Salmonella enterica serovar Typhi isolated from Bangladesh

Emilie Westeel * [1], Arif Tanmoy [2], Nicholas Lima [3], Alain Rajoharison , Katrien De Bruyne [4], Johan Goris [4], Luiz Gonzala [3], Alex Van Belkum [4], Ana Tereza R. Vasconcelos [5], Samir Saha [2], Hubert Endtz , Florence Komurian-Pradel

[1] Fondation Mérieux - Laboratoire des Pathogènes Emergents (LPE) – CIRI, Centre Internatinal de Recherche en Infectiologie – France
[2] Child Health Research Foundation (CHRF) – Bangladesh
[3] National Laboratory for Scientific Computing (LNCC) – Brazil
[4] Biomérieux – Belgium
[5] Laboratorio Nacional de Computação Cientifica / National Laboratory for Scientific Computation (LNCC / MCT) – LNCC, Av. Getulio Vargas, 333, Quitandinha, 25651-075, Petropolis, RJ, Brazil

Typhoid fever, caused by *Salmonella enterica* serovar Typhi (*S.* Typhi) has become a major public health concern globally, due to increasing antimicrobial resistance (AMR) and shrinking list of treatment options. Disease-endemic countries like Bangladesh, require detail genetic characterization of AMR to fight against.
Here, we studied whole genome sequencing (WGS) data of 545 *S.* Typhi isolates, collected during 1999-2013, from Bangladesh. We found high sensitivity and specificity of WGS while predicting the AMR phenotypes for ampicillin (amp), chloramphenicol (chl), cotrimoxazole (sxt) and ceftriaxone (cro), while ciprofloxacin (cip) needs further adjustment. Genotype 4.3.1, usually associated to multidrug resistance (MDR), was dominant in the country. *bla*CTX-M-15 gene was detected to cause ceftriaxone resistance, an antibiotic belonging to the third-line generation, same as the recent outbreak in Pakistan, except the genotype and phenotypes were different.

Multidrug resistance (amp,sxt,chl) in *S.* Typhi is largely described as being encoded on IncHI1 plasmids. However, only 47 strains among the 208 MDR strains present in our samples contain an IncHI1 plasmid. The remaining 161 strains just contain parts (6-12%) of IncHI1, corresponding to resistance genes and insertion sequences. In 2015, a genomic island named SGI11 (25kb) have been discovered in MDR *S.* Typhi strain isolated from Bangladesh. This island carries 7 resistance genes (blaTEM-1, catA1, strA, strB, sul1, sul2, and dfrA7). Globally, among our strains, the IncHI1 plasmid seems to be lost with time, in favor of genomic island integration.

To go further into the analyses, 73 isolates have been selected among the 545, according to their antimicrobial resistance profiles and submitted to complete genome closure, annotation, and comparative genomics. Genetic elements responsible for AMR (e.g. genes, mutations and, genomic islands) and their genomic location (plasmid or, chromosome) were analyzed. Exact

[*] Speaker

location of the genomics islands detected have been determined. We identified 5 SGI11 variation, differing by their resistance gene content, so conferring different resistant phenotypes (not only MDR). We also detected 5 types of plasmids, with 3 of them carrying resistance genes.
WGS analysis combined with clinical metadata provides insight on resistance mechanisms in *Salmonella* Typhi, and could be used for adapting antibiotic treatment regimes.

Keywords: WGS, salmonella typhi, antimicrobial resistance

Pan-genomic analysis to redefine species and subspecies

Aurélia Caputo * [1], Didier Raoult [2]

[1] Institut Hospitalier Universitaire Méditerranée Infection (IHU) – MEPHI, AP-HM, Aix-Marseille Université - AMU – 19-21 Bd Jean Moulin 13385 MARSEILLE Cedex 05 FRANCE, France
[2] Institut Hospitalier Universitaire Méditerranée Infection (IHU) – MEPHI – 19-21 Bd Jean Moulin 13385 MARSEILLE Cedex 05 FRANCE, France

Since the introduction of DNA sequencing by Sanger and Coulson in 1977, considerable progress has been made. A growing number of data is being generated in several areas and requires more and more advances in computing. Bio-informatics is essential today in many fields such as data management and analysis, genomics with assembly and genome annotation, comparative genomics, phylogeny, metagenomics, research new bacterial species and taxonomic classification. Taxonomy is a set of many changes based on available data, methods used and evolution of bacterial identification techniques.
Various methods are currently used to define species and are based on the phylogenetic marker 16S ribosomal RNA gene sequence, DNA-DNA hybridization and DNA GC content. However, these are restricted genetic tools and showed significant limitations.

We describe an alternative method to build taxonomy by analyzing the pan-genome composition of different species of the *Klebsiella* genus. *Klebsiella* species are Gram-negative bacilli belonging to the large *Enterobacteriaceae* family. Interestingly, when comparing the core/pan-genome ratio; we found a clear discontinuous variation that can define a new species.
Using this pan-genomic approach, we showed that *Klebsiella pneumoniae* subsp. *ozaenae* and *Klebsiella pneumoniae* subsp. *rhinoscleromatis* are species of the *Klebsiella* genus, rather than subspecies of *Klebsiella pneumoniae.* This pan-genomic analysis, helped to develop a new tool for defining species introducing a quantic perspective for taxonomy.

Keywords: genomics, pan, genome, species definition, Klebsiella pneumoniae, taxonomy

*Speaker

ToulligQC: A MinION run data analysis tool

Berengere Laffay *† [1,2], Ammara Mohammad [1], Corinne Blugeon [1], Fanny Coulpier [1], Stéphane Le Crom [1,3], Sophie Lemoine [1], Laurent Jourdren [1]

[1] Institut de biologie de l'école normale supérieure (IBENS) – École normale supérieure [ENS] - Paris, Inserm : U1024, CNRS : UMR8197 – 46, Rue d'Ulm. 75005 Paris, France
[2] Master Bioinformatique, Normandie Université, UNIROUEN (UNIROUEN) – Université de Rouen Normandie – UNIVERSITÉ DE ROUEN NORMANDIE Place Émile Blondel - 76821 Mont-Saint-Aignan cedex, France
[3] Institut de Biologie Paris Seine – Sorbonne Université UPMC Paris VI – France

The MinION device developed by Oxford Nanopore Technologies (ONT) is a portable device that aims to produce long DNA sequence (up to 200 kb) and full-length RNA. Illumina sequencers only produce short reads (100 to 300 bp).
The Illumina technology sequencing principle is based on synthesis using a proprietary reversible terminator-based method. The sequencer will detect single bases as they are incorporated into template strands (the emission wavelength and intensity are used to identify the sequence).

The ONT sequencing principle is rather different: the sequencing process will occur when a DNA or a RNA sequence goes through a nanopore. The ionic current passing through the nanopores changes as a k-mer of nucleotides passes. This current modification is specific of a k-mer and can be used to define the sequence when the basecalling is performed.

MinKNOW carries out the data acquisition during the run and produces Fast5 files using the HDF5 format to store the data. The electrical signal is then translated into a nucleic acid sequence using Albacore, the ONT basecaller. In the end, the sequence and the metadata are written a Fast5 file.

The existing run data analysis tools developed for Illumina runs are neither adapted to the Fast5 file format nor to the quality metrics that suit long reads. MinKNOW provides metrics and scales that are not appropriated to RNA sequences. It was necessary to develop an QC tool dedicated to ONT runs and flexible enough to handle DNA and RNA sequencing.

This poster presents ToulligQC, a program dedicated to ONT run data analyses, RNA or DNA. ToulligQC provides a detailed HTML report of quality metrics like read lengths and Phred scores through a set of graphs.

It is possible to handle multiple samples on an ONT run, each sample being identified by its barcode (added during the library preparation). ToulligQC then allows to retrieve each sample metrics and distribution at the end of the run.

*Speaker

†Corresponding author: laffay@biologie.ens.fr

ONT protocols are in constant development, so ToulligQC has to be flexible enough to be adapted quickly. It now supports the latest version of Albacore (2.X) and can be used to handle the different types of sequencing, 1D and 1D2, proposed by ONT.

It is an **open source** software which can be freely downloaded on **Github** [1], as a **Docker image** (genomiqueparisc entre/toulligqc), and as a **Pypy package** [2].

Bibliography :

https://github.com/GenomicParisCentre/toulligQC
https://pypi.python.org/pypi/toulligqc/0.6

Keywords: MinION, Oxford Nanopore, pipeline, analysis, QC

An evaluation of Oxford Nanopore data in RNASeq projects

Berengere Laffay *† [1,2], Stéphane Le Crom‡ [2,3], Laurent Jourdren * § [2], Sophie Lemoine * ¶ [2]

[1] Master Bioinformatique, Normandie Université, UNIROUEN (UNIROUEN) – Université de Rouen Normandie – UNIVERSITÉ DE ROUEN NORMANDIE Place Émile Blondel - 76821 Mont-Saint-Aignan cedex, France
[2] Institut de biologie de l'école normale supérieure (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris, CNRS : UMR8196, Institut National de la Santé et de la Recherche Médicale - INSERM – France
[3] Institut de Biologie Paris Seine (EPS-IBPS) – Sorbonne Université UPMC Paris VI – France

We have been producing and analyzing long-reads using a MinION sequencer from Oxford Nanopore Technology (ONT) for two years. As a functional genomics core facility [1], our main concern is RNASeq and ONT long-reads are a good opportunity to decipher RNA isoform usage as one read can cover a whole transcript.
The goal of this poster is to give an overview of the data we have been analyzing, their defects and qualities, what can be expected from ONT long-reads and what needs better developments in a RNA environment.

As an evaluation, we sequence RNAs from the same validation design (3 x Egr2 KO mice versus 3x WT mice) each time we want to test a new protocol on a technology enhancement. We therefore have a huge amount of data to be compared. This RNASeq design was performed on Illumina sequencers and protocols but also on the ONT protocols. We can now compare the sequences and alignments of:

- short-reads versus ONT long-reads from cDNA libraries,

- ONT 1D cDNA data versus 1D2 cDNA data,

- ONT Direct RNA versus ONT 1D cDNA data.

Most of the runs were performed 3 times. We are not yet able to point out what biases have to be taken into account to ensure a good data analysis: the flowcells, the protocols, the algorithms evolve too quickly to define a good statistical model. If we cannot talk about statistics, we can roughly evaluate the reproducibility of what we have done so far.

As second generation sequencing has been on the scene for 10 years, most of the tools to analyze

*Speaker
†Corresponding author: laffay@biologie.ens.fr
‡Corresponding author: lecrom@biologie.ens.fr
§Corresponding author: jourdren@biologie.ens.fr
¶Corresponding author: slemoine@biologie.ens.fr

sequencing data are dedicated to short-reads and cannot be used for long-reads. We developed ToulligQC [2] to collect RNASeq run data and evaluate its quality. Routines to perform secondary data QC and then further analysis are not gold standards. This poster we make a short summary of the tool dimension.

Bibliography :

http://genomique.biologie.ens.fr
https://github.com/GenomicParisCentre/toulligQC

Keywords: MinION, Oxford Nanopore Technology, Long reads, RNASeq

Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening

Ophélie Jouffroy * , Surya Saha , Lukas Mueller , Hadi Quesneville [1], Florian Maumus [2]

[1] Unité de Recherche Génomique Info (URGI) – Institut national de la recherche agronomique (INRA) : UR1164 – INRA, Centre de recherche de Versailles, bat.18 Route de Saint Cyr 78000 Versailles, France
[2] URGI – Institut National de la Recherche Agronomique - INRA (FRANCE) – France

Background: Plant genomes are populated by different types of repetitive elements including transposable elements (TEs) and simple sequence repeats (SSRs) that can have a strong impact on genome size and dynamic as well as on the regulation of gene transcription. At least two-thirds of the tomato genome is composed of repeats. While their bulk impact on genome organization has been recently revealed by whole genome assembly, their influence on tomato biology and phenotype remains largely unaddressed. More specifically, the effects and roles of DNA repeats on the maturation of fleshy fruits, which is a complex process of key agro-economic interest, still needs to be investigated comprehensively and tomato is arguably an excellent model for such study.

Results: We have performed a comprehensive annotation of the tomato repeatome to explore its potential impact on tomato genome composition and gene transcription. Our results show that the tomato genome can be fractioned into three compartments with different gene and repeat density, each compartment presenting contrasting repeat and gene composition, repeat-gene associations and different gene transcriptional levels. In the context of fruit ripening, we found that repeats are present in the majority of differentially methylated regions (DMRs) and thousands of repeat-associated DMRs are found in gene proximity including hundreds that are differentially regulated. Furthermore, we found that repeats are also present in the proximity of binding sites of the key ripening protein RIN. We also observed that some repeat families are present at unexpected high frequency in the proximity of genes that are differentially expressed during tomato ripening.

*Speaker

Keywords: Fruit ripening, DNA methylation, Transposable elements, Tomato

A tool to genotype individuals and assemble allelic sequences at highly polymorphic loci from raw NGS reads data: application to the self-incompatibility locus of Brassicaceae

Mathieu Genete *† [1], Sophie Gallina [1], Vincent Castric [1], Xavier Vekemans [1]

[1] Laboratoire Evolution, Ecologie et Paléontologie (EEP) – Université Lille I - Sciences et technologies, CNRS : UMR8198 – France

Loci with extremely high levels of molecular polymorphism such as the self-incompatibility locus (S-locus) of Brassicaceae have remained recalcitrant to genotyping with NGS technologies based on short reads, as they are typically challenging to assemble de novo as well as to align to a given reference. Up to now, studies of the allelic diversity at the S-locus in natural populations have relied on labor-intensive molecular cloning or BAC library approaches. Due to the severe reduction of the cost of shotgun sequencing, obtaining raw reads from individual genomes is now becoming possible on a large scale and our previous work has shown that such data can be used to reliably genotype individual accessions from the 1001 genome project of Arabidopsis thaliana at the S-locus (Tsuchimatsu et al. 2017, https://doi.org/10.1093/molbev/msx122). Here, we present an efficient pipeline to map raw reads from individual outcrossing Arabidopsis genomes against a dataset of multiple reference sequences of the pistil specificity determining gene of the Brassicaceae S-locus (SRK) and determine individual S-genotypes. In line with the important trans-specific polymorphism observed in this genetic system, we show that this approach can be first used to successfully obtain S-locus genotypes in related Brassicaceae genera, even if the species is not included in the initial database. We further show that this approach can be used to specifically assemble full-length individual S-allele sequences, and even discover new allelic sequences that were not initially present in the database. This pipeline can in principle be adapted to other highly polymorphic loci, given datasets of reference sequences are available. The pipeline will be available from a Docker Hub repository, as a docker file or image which contains all third-party tools for immediate use.

Keywords: genotyping, assembly, allele, loci, polymorphism, NGS, illumina

*Speaker

†Corresponding author: mathieu.genete@univ-lille1.fr

Extraction of biologically meaningful patterns from high-dimensional omics data, a platform based on Self-Organizing Maps

Shingo Miyauchi [1], Marie-Noëlle Rosso * [1]

[1] Biodiversité et Biotechnologie Fongiques (BBF) – Institut national de la recherche agronomique (INRA), Aix-Marseille Université - AMU – UMR 1163, Polytech Marseille, 163 Avenue de Luminy, CP925, 13288 Marseille, Cedex 9, France

Genomics, transcriptomics and proteomics can be used individually to address the genes, transcripts and proteins related to biological functions. Overlaying the three levels of information sharpens our understanding and highlights different levels of gene expression regulations. However, genome-wide transcriptomic and proteomic activities are complex. For example, capturing just a single time point of fungal transcriptomic activity involves more than ten thousand genes showing various transcription levels. The number of observations increases exponentially when we add the number of biological replicates, different growth conditions, and time points. The addition of proteomic information gives an extra layer of complexity.
To extract biologically meaningful patterns from such high-dimensional omics data, we have developed a multi-omics profiling platform, Self-organizing map Harboring Informative Nodes with Gene Ontology (SHIN+GO). Genome-wide omics models constructed with the platform are designed to pinpoint biological activities of interest that would otherwise be buried in the high-dimensional data. One of the key components of this platform, Self-organizing map (SOM) is an algorithm constructing a neural network with given input data in an unsupervised manner. SOM reduces the number of features in high-dimensional data by grouping similar items and forming clusters. It has a unique property of making two-dimensional maps suitable for large-scale data visualization.

The first part of the SHIN+GO platform, Self-organizing map Harboring Informative Nodes (SHIN) generates neural networks of genome-wide transcript levels and highlights condition-specific responses in transcriptomes Next, the count of secreted proteins is overlaid onto the master SOM. As a result of this integration of data, SHIN provides nodes made of clustered co-regulated genes (transcriptomes) with corresponding co-secreted proteins (secretomes).

The second part of the SHIN+GO platform, Gene Ontology (GO), was developed to; 1) measure the frequency of gene functional annotations present in the nodes; and 2) biologically interpret the outputs of the genome-wide omics models generated. Biological terms with statistically enriched occurrence in a node were used as an indicator of biological functions for the grouped genes and proteins.

In this study, we used the platform (SHIN+GO) to generate dynamic genomewide integrative

*Speaker

omics models with a recently-sequenced fungus, *Pycnoporus coccineus*, to establish transcriptomic and secretomic profiles during plant biomass decomposition. We used the models as a guide to capture biologically interesting omics hotspots related to the fungal adaptive responses to different plant biomasses.

The method we describe is versatile and can be used for large omics data from any genome-sequenced organisms.
Reference: Miyauchi S, Navarro D, Grisel S, Chevret D, Berrin J-G, Rosso M-N (2017) The integrative omics of white-rot fungus *Pycnoporus coccineus* reveals co-regulated CAZymes for orchestrated lignocellulose breakdown. PLoS ONE 12(4): e0175528. https://doi.org/10.1371/journal.pone.0175528

Keywords: omics, genomics, transcriptomics, secretomics, integrative, map

Deciphering the origins of enzymes substrate specificity using large-scale sequence analyses

Clothilde Chenal *† , Ludovic Pelosi , Fabien Pierrel , Sophie Abby [1], Ivan Junier

[1] Techniques de lÍngénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications [Grenoble] (TIMC-IMAG) – Centre National de la Recherche Scientifique : UMR5525, Université Grenoble Alpes – Domaine de la Merci - 38706 La Tronche, France

Predicting the molecular function of an enzyme from its sequence alone is challenging for multiple reasons.

One common difficulty, for instance, lies in the fact that slight variations in the sequence of an enzyme can have deep impacts on its substrate specificity. A case in point concerns enzymes known as flavin-containing monooxygenases (FMO) and involved in the biosynthesis of ubiquinone, a key molecule in the respiratory chain of proteobacteria and mitochondria. Namely, the synthesis of ubiquinone involves the hydroxylation of three positions of an aromatic ring, which are implemented by three different FMOs in Escherichia coli. In other distant bacteria, though, FMOs have been shown to hydroxylate two or even three positions of the aromatic ring and have therefore a broad regioselectivity (see Figure). Thus, in these organisms fewer enzymes are involved in the ubiquinone pathway [1]. These observations raise two fundamental questions: i) what are the mechanisms dictating the degree of regioselectivity of FMOs? and ii) is it possible to identify sequence features associated to these mechanisms?

In our work, we aim at answering these questions by leveraging the large number of fully sequenced genomes available in public databases. To this end, we use a comparative genomics approach and confront clustering properties of FMOs in sequence space with the phylogenetic profiling of these enzymes among a thousand genomes. We also perform amino acid co-evolution analyses to go beyond the identification of sequence motifs and identify global pattern of cooperativity between amino acids that may underlie the variation of regioselectivity.

As a result, we have identified patterns that we predict to dictate the regioselectivity of certain

*Speaker

†Corresponding author: clothilde.chenal@etu.umontpellier.fr

FMOs and we are
currently testing our predictions experimentally. Our work also opens the way to a fine annotation of closely related
enzymes for which standard approaches are poorly adapted, as in the case of the discrimination of paralogs.

References :

Pelosi, L., Ducluzeau, A.-L., Loiseau, L., Barras, F., Schneider, D., Junier, I., Pierrel, F. (2016). Evolution
of Ubiquinone Biosynthesis: Multiple Proteobacterial Enzymes with Various Regioselectivities To Catalyze
Three Contiguous Aromatic Hydroxylation Reactions. mSystems, 1(4), e00091–16–16.

Keywords: sequence motif analysis, molecular function, amino acid coevolution, enzyme substrate specificity, regioselectivity, ubiquinone

Genetic diversity of staphylococcal strains isolated from food and enterotoxin coding genes

Arnaud Felten * [1], Déborah Merda , Noémie Vingadassalon , Michel-Yves Mistou , Jacques-Antoine Hennekinne

[1] ANSES – Anses – France

Introduction: Identify and characterize pathogens responsible for a food outbreak is necessary in public health and quality control in industries. Some pathogenic strains belonging to the bacterial species, *Staphylococcus aureus* can produce toxins in food, which could lead to staphylococcal food poisoning outbreaks (SFPO).
Purpose: Objectives of the study were to use whole genome sequencing to determine the genetic diversity of strains responsible for SFPO to 2005 to 2017 in Europe and the most frequent enterotoxin genes in these strains.

Materiel and methods: A collection of 143 genomes was sequenced using illumina Technology. This collection was composed of strains responsible for SFPO and of reference strains isolated from food, environment or humans. In order to study genetic diversity of *S. aureus* strains isolated from food within known genetic diversity of *S. aureus*, 105 genomes available from public database were included. Assembly and annotation were performed using in-house workflow based on Spades and Prokka. The core genome was defined using roary, and the phylogeny was performed using RAxML. Then, toxinic profiles were established on the 23 enterotoxin genes available in the literature by using an in-house workflow based on blast approach. Finally the genetic diversity of enterotoxin coding genes was studied using clustering approaches.

Results: Our results allowed to highlight several divergent clones within *S. aureus* were responsible for SFPO between 2005 and 2017. Furthermore, several enterotoxin coding genes were very frequents in these strains, as *seg*, *seh*, *sem*, *sen* and *seo*. Finally, clustering of different alleles of enterotoxin coding genes showed a genetic proximity between *sea*, *sep* and *see* genes.
Significance: These results are relevant for food safety as they allowed us i) to highlight the presence of enterotoxin coding genes not currently detected by PCR tools and ii) to determine new targets for the development of rapid detection methods.

Keywords: Staphylococcus aureus, food outbreak, toxins, genetic diversity

*Speaker

A Multiplex Network approach to Premature Aging Diseases

Alberto Valdeolivas *† [1,2], Claire Navarro [1], Sophie Perrin [1], Pierre Cau [1,2], Anaïs Baudot‡ [1]

[1] Marseille medical genetics - Centre de génétique médicale de Marseille (MMG) – Aix Marseille Université : UMR$_S$1251, $InstitutNationaldelaSantéetdelaRechercheMédicale$: $UMR_S1251 - -FacultédeMédecine - Timone27, boulevardJeanMoulin13385Marseillecedex5, France$

[2] ProGeLife – ProGeLife – 8 Rue Sainte Barbe 13001, Marseille, France

Premature aging (PA) syndromes are a group of heterogeneous rare disorders that recapitulate some of the aspects associated to physiological aging. They are caused by mutations in several genes involved in different biological processes. Genes and proteins do not act isolated in cells but rather interact in complex networks of molecular interactions. In this context, we undertook a network approach to better understand the etiology and pathophysiolgy of these diseases.

First, we extracted the network modules surrounding genes mutated in PA diseases, to define the landscape of biological processes that might be perturbed. To this goal, we applied a strategy based on our recently developed random walk (RW) with restart on multiplex networks [1]. This allows us to navigate and extract information from different layers of physical and functional interactions (e.g., protein-protein, co-expression, molecular complexes) outperforming single-network approaches [1]. We captured modules representing the hallmarks of physiological aging, and compared the processes commonly perturbed in PA diseases, as well as those specific to a subset of diseases.

In a second part, we are developing a strategy to analyse the impact on networks of PA disease-causing mutations. To this goal, we are performing targeted attacks, removing from the multiplex network either genes (to simulate loss-of-function) or some of their interactions (to simulate ”edgetic” mutations). A modified version of our RW algorithm allows us to study the topological modifications of the network after the attack, pinpointing to the most affected genes, modules and processes.

Valdeolivas,A. et al. Random Walk With Restart On Multiplex And Heterogeneous Biological Networks. 2017. bioRxiv.

Keywords: Aging, Rare Diseases, Premature Aging, Networks, Multiplex Networks, Modules, Ran-

[*] Speaker

[†] Corresponding author: alvaldeolivas@gmail.com

[‡] Corresponding author: anais.baudot@univ-amu.fr

dom Walk, Layers, Network Attacks, Edgetic

Integrative visual omics of the white-rot fungus Polyporus brumalis exposes the biotechnological potential of its oxidative enzymes for delignifying raw plant biomass.

Elodie Drula * [1], Shingo Miyauchi [1], Anaïs Rancon [2], Bernard Henrissat [3], Matthieu Heinaut [2], Francisco Ruiz-Dueñas [4], Isabelle Herpoël-Gimbert [5], David Navarro [1], Igor V. Grigoriev [6], Simeng Zhou [7], Sana Raouche [5], Marie-Noelle Rosso [1]

[1] INRA – Institut National de la Recherche Agronomique – France
[2] n/a – Aucune – France
[3] CNRS – Centre National de la Recherche Scientifique - CNRS – France
[4] Centro de Investigaciones Biologicas – Spain
[5] Aix Marseille Université (AMU) – Aix-Marseille Université - AMU – Aix-Marseille UniversitéJardins du Pharo58 Boulevard Charles Livon13284 Marseille cedex 7, France
[6] University of California – United States
[7] Institut des Sciences Moléculaires de Marseille (ISM2) – Aix Marseille Université : UMR7313, Ecole Centrale de Marseille : UMR7313, Centre National de la Recherche Scientifique : UMR7313 – Campus Saint Jérôme Av. escadrille Normandie Niemen BP 531 13397 MARSEILLE CEDEX 20, France

Plant biomass conversion for green chemistry and bio-energy is a current challenge for a modern sustainable bioeconomy. The complex polyaromatic lignin polymers in raw biomass feedstocks (i.e. agriculture and forestry by-products) are major obstacles for biomass conversions.
White-rot fungi can degrade all plant cell wall polymers through the concerted secretion of complex sets of hydrolytic and oxidative enzymes. These enzymes belong to enzyme families including glycoside hydrolases (GH), carbohydrate esterases (CE), pectate lyases (PL), and auxiliary oxido-reductases (AA) as classified in the Carbohydrate Active Enzyme database [CAZy; www.CAZy.org; (1)]. In particular, the degradation of crystalline cellulose is facilitated by cellobiohydrolases (GH6 and GH7) and lytic polysaccharide monooxygenases (LPMOs; CAZy family AA9), which are often linked to Carbohydrate Binding Modules (CBM1). In addition, genes coding for Class II peroxidases of the peroxidase-catalase superfamily involved in the oxidative breakdown of lignin [(2); CAZy family AA2] are a hallmark of white-rot fungi.

The white-rot fungus Polyporus brumalis efficiently breaks down lignin and is regarded as having a high potential for the initial treatment of plant biomass in its conversion to bio-energy.

The goal of our study was to understand the lignin degrading capability of P. brumalis during growth on wheat straw, a lignocellulosic substrate that is considered as a biomass feedstock worldwide.

*Speaker

We inspected whether CAZy gene family expansions occurred in the genome of P. brumalis using CAFE, a computational tool that provides statistical analysis of evolutionary changes in gene family size over a phylogenetic tree (3). We observed the co-occurrence of gene family expansions for putatively secreted lignin-active peroxidases and H2O2-generating enzymes, which could contribute to the distinctive ability of P. brumalis for selective delignification of raw biomass.

We conducted an integrative multi-omics analysis by combining data from the fungal genome, transcriptomes, and secretomes. We used the visual multi-omics pipeline SHIN+GO to identify co-regulated genes showing similar transcription patterns throughout the culture on wheat straw by first integrating time-course transcriptomes with corresponding co-secreted proteins and then converting these data into genome-wide graphical network maps (4). These omics topographies ('Tatami maps') allowed us to :

1) visualize nodes containing genes with similar transcription patterns with the corresponding count of secreted proteins;

2) calculate the node-wise mean of the normalized transcript read counts in each condition;

3) identify gene clusters showing high transcription levels at specific time points and under specific conditions.

The examination of interrelated multi-omics patterns revealed the coordinated regulation of lignin-active peroxidases and H2O2 -generating enzymes along with the activation of cellular mechanisms for detoxification, which combined to result in the efficient lignin breakdown by the fungus.

References

1.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho P, Henrissat B. 2014. The carbohydrate-active enzymes database (cazy) in 2013. Nucleic Acids Res 42:D490-495.

2.

Zámocký M, Hofbauer S, Schaffner I, Gasselhuber B, Nicolussi A, Soudi M, Pirker KF, Furtm´uller PG, Obinger C. 2015. Independent evolution of four heme peroxidase superfamilies. Arch Biochem Biophys 574:108-119.

3.

Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using cafe 3. Mol Biol Evol 30:1987-1997.

4.

Miyauchi S, Navarro D, Grigoriev IV, Lipzen A, Riley R, Chevret D, Grisel S, Berrin

J-G, Henrissat B, Rosso M-N. 2016. Visual comparative omics of fungi for plant biomass deconstruction. Front Microbiol 7:1335.

Keywords: Fungal adaptive response to lignocellulose

Context-specific prioritization of non-coding variants implicated in human diseases

Lambert Moyon *† [1], Camille Berthelot [1], Hugues Roest-Crollius‡ [1]

[1] Institut de biologie de lÉNS Paris (UMR 8197/1024) (IBENS) – École normale supérieure - Paris, Institut National de la Santé et de la Recherche Médicale : U1024, Centre National de la Recherche Scientifique : UMR8197 – 46, Rue d'Ulm75005 Paris, France

Whole genome sequencing is increasingly being used for patients with rare genetic diseases as a diagnostic tool. However, for a large proportion of sequenced patients, no coding mutation is found in a gene associated with the phenotype. In these cases, a non-coding mutation, located in a cis-regulatory region, may affect the expression of a gene involved in the disease. Despite the existence of methods for annotating and predicting regulatory sequences on the basis of biochemical and epigenetic properties, defining objective criteria remains difficult to effectively select candidates among the millions of non-coding mutations present in each patient. Moreover, the mechanisms of action and interaction between regulatory regions and target genes are still unclear, making it difficult to link a non-coding mutation with the patient's phenotype.

We propose here a supervised machine learning strategy using random forests, adapted to complex and heterogeneous datasets, to classify and select non-coding mutations potentially involved in the deregulation of disease genes. A notable innovation of our approach is to take into account association data between non-coding regions and target genes.

We apply 3 classifiers, trained on different sets of experimentally predicted regulatory regions, on more than 40,000 non-coding mutations in 48 patients affected with X-linked intellectual disabilities from the FP7-funded project " NeuroXsys ". Selected mutations were shown to segregate with the disease in affected families, and to deregulate the predicted target gene in animal models. We discuss the results in light of their genome-wide application to larger cohorts of patients.

Keywords: machine Learning, non, coding variants, whole, genome sequencing, disease, regulation, enhancers

*Speaker

†Corresponding author: moyon@biologie.ens.fr

‡Corresponding author: hrc@ens.fr

Characterization of Escherichia coli K-12 regulatory networks by bioinformatics integration of high-throughput data

Claire Rioualen * [1], Julio Collado Vides [2], Jacques Van Helden [1]

[1] Technologies avancées pour le génôme et la clinique (TAGC) – Inserm : U1090, Université de la Méditerranée - Aix-Marseille II – Parc scientifique de Luminy - 163 avenue de Luminy 13288 Marseille cedex 9, France

[2] Centro de Ciencias Genómicas (CCG) – av. de la Universidad col. Chamilpa 62210 Cuernavaca, Mor., Mexico

Background

Gene regulation is essential to any living organism, whether in physiological conditions or in response to environmental stimuli, and involves a variety of mechanisms. One of the most common ones is the binding of transcription factors on specific sites of the DNA, called TFBS. By interacting with the recruitment of the RNA polymerase complex, it can have an impact on the transcription of the surrounding genes, whether positive or negative.

Those mechanisms can be characterized by using Next-Generation Sequencing (NGS) technologies. ChIP-seq1,2 allows to characterize DNA binding locations of transcription factors (TFs) at a genomic scale, using a reference genome. RNA-seq technology, or whole transcriptome shotgun sequencing, allows to quantify transcription of all the genes in a given cell, and compare the levels of transcription between different conditions.

Problematics

Escherichia coli K-12 is a model organism particularly adapted to the study of gene regulation mechanisms, and its genome is already well characterized. A milestone in the study of regulatory mechanisms was the description of the Lac operon using this strain (Jacob and Monod, 1961). Its genome was one of the first to be entirely sequenced and published (Blattner et al., 1997). Extensive information about TFs, their binding sites, target genes and operons has been manually curated and indexed for more than 20 years in dedicated databases such as RegulonDB (Gama-Castro et al., 2016) and EcoCyc (Keseler et al., 2017).

However, so far NGS technologies have barely been applied to E.coli K-12, nor bacteria at large, and no guidelines exist on how to analyse such data in prokaryotic organisms. Furthermore, the existing tools for NGS data analysis were mostly developed for eukaryotic genomes, which have different characteristics of genome size and organization.

Methods

*Speaker

We have developed Snakemake-based workflows (Koster et al., 2012) as part of the SnakeChunks project (Rioualen et al., submitted), in order to analyse ChIP-seq and RNA-seq data targeting the same TFs in E.coli K-12. The modular structure of the workflows allows to use a variety of tools and parameters for each step of the analyses. It can thus be used to perform benchmarking of the crucial steps that are the peak-calling and the detection of differentially expressed genes. It has been shown in human datasets that the choice of tools has a great impact on the results (Pepke et al., 2009; Bailey et al., 2013), and most tools were primarily designed for the analysis of eukaryotic data. No such work has been realized in bacteria so far.

Once peak-calling and differential gene expression analysis are properly performed, a lot of information can be extracted or deduced from the resulting data. TFs are characterized by their binding sites (TFBS) and target genes. However, most of E.coli's TFs are poorly characterized: about a hundred of them don't have any known binding sites, and another hundred have between one and three TFBSs. By using ChIP-seq data, many new TFBSs can be annotated genome-wide. Furthermore, these TFBSs allow to build weight matrices and thus associate binding motifs to each TF. The localization of TFBSs also allows to infer hypotheses about potential target genes. This can be confirmed by integration of RNA-seq data and analysis of differential expression. However, the association of ChIP-seq peaks and differentially-expressed genes is not a straightforward process, and many factors can interfere with the regulation of a target gene by its TF (Myers et al., 2013). In such cases, performing motifs discovery can unravel interactions with other TFs, whether cooperative or competitive. Finally, RNA-seq data allows to identify new transcription units (TUs), and potentially identify new operons altogether. These TUs and operons can be associated with annotated TSSs, as well as predicted ones (Mendoza-Vargas et al., 2009; Thomason et al., 2015).

Results

Preliminary results were obtained by reanalyzing a dataset combining ChIP-seq and RNA-seq and aimed at characterizing the Fumarate and Nitrate reductase Regulatory (FNR) transcription factor genome-wide (Myers et al., 2013). The analysis of RNA-seq data revealed that the expression of 484 genes was impacted by the deletion of FNR, which is consistent with the fact that this TF regulates many essential processes such as anaerobic growth and acid resistance in E.coli. However, the exhaustive curation of previously published experiments targeting FNR, available in RegulonDB (Gama-Castro et al., 2016), revealed that about 75% of the discovered genes had not yet been identified as FNR target genes, and 12% of them were corroborated by the ChIP-seq data. This is the case for the leuLABCD operon, for instance, which is known mainly for encoding the enzymes responsible for the biosynthesis of leucine from valine. Despite not being listed as a target of FNR, its genes show a lower expression in the FNR mutant, suggesting that FNR could potentially regulate its expression. This hypothesis is enforced by the ChIP-seq profiles, since we can observe a clear peak upstream of the operon. This interpretation relying on the integrated results of SnakeChunks workflows is consistent with Myers and colleagues' observations. This work was submitted to Current Protocols in Bioinformatics (Rioualen et al., under reviewing).

Perspectives

Until today, most knowledge of regulation in E.coli K-12 has been accumulated through low-throughput experiments, and has been manually curated in RegulonDB (Gama-Castro et al., 2016). High-throughput data hasn't been integrated yet, and though there are few datasets available at the moment, the amount is growing exponentially, which is why developing modular and reproducible workflows for the analysis of such data is becoming a necessity. It will allow, in a near future, to perform automated biocuration of any newly published data.

The next goal of this project is to make these workflows and tools accessible to all researchers, whether bioinformaticians, experimentalists, or biocurators, and allow them to analyse their own data and customize their choice of tools and parameters.

The combination of these large-scale experiments allows to answer many questions on gene regulatory mechanisms, and further characterizing E.coli would help understand its amazing adaptability, as well as explain why certains strains are pathogenic.

Keywords: gene regulation, Escherichia coli K, 12, ChIPseq, RNAseq, NGS, model organism, reproducibility

Knowledge management and standard representation of causal statements: new resources for systems modelling

Vasundra Touré *† [1], åsmund Flobak [1], Astrid Lægreid [1], Martin Kuiper [1]

[1] Norwegian University of Science and Technology [Trondheim] (NTNU) – NO-7491 Trondheim, Norway

In Systems Biology, regulatory process networks are built to reflect how components in cell fate decision systems are interconnected and behave. A considerable amount of knowledge provided by different public resources is available in the form of large biological networks depicting metabolic reactions, signaling cascades or even gene regulatory events. We aim at exploring and using this information by breaking down those networks into their most basic regulatory network motifs, called "causal statements". A causal statement is characterised by a directed interaction where a source entity (regulator) has an influence over the quantity or the activity of a target entity (regulated). By looking at the core interactions occurring among entities, the understanding of the mechanisms they enable in biological regulations could be improved.

The DrugLogics project (https://www.ntnu.edu/health/druglogics) is a Systems Medicine approach to employ computational methods for predicting drug resistance in cancer treatment. The long-term goal is to economize drug screens and to find tailor-made treatments for patients with specific types of cancer. Today we have a pipeline that automatically generates boolean models from a repository of causal statements to predict the effect of drugs and drug combinations. Our contribution consists of 1) building strategies to extract causal statements from existing network resources to feed our model building software pipeline with a comprehensive set of causal interactions, 2) standardising the representation of causal statements and the logical models generated to improve the information content and the intelligibility, and the reusability of the produced data.

We will design a format to standardize the representation of causal statements by using generally accepted identifiers and ontology terms. In order to facilitate this task, we are establishing curation guidelines, called the "Minimal Information about a Causal Statement" (MICAST) in collaboration with members from the International Molecular Exchange Consortium (IMEx). The guideline (draft available at: https://github.com/vtoure/MICAST) will provide recommendations on information to depict and ontologies to use when delineating a causal statement. Based on this, the formal representation will provide contextual information (e.g, species, cell type, experimental setup) to facilitate the implementation and exploration of causal data and avoid an ambiguous description of it. With this representation format we will design and produce software pipelines to extract causal statements from a variety of existing network resources such as Reactome, the Atlas of Cancer Signaling Network, etc... As some databases may not explicitly provide causal statements or structure their data as binary causal interactions, we

*Speaker

†Corresponding author: vasundra.toure@ntnu.no

are building strategies to translate reaction networks into causal interactions. This involves an inference of causation based on the biological motifs found and the necessity of adding contextual information in the causal data retrieved. Finally, the repository of causal statements generated aims to be a valuable public knowledge resource that would facilitate the process of model building.

Keywords: causal statement, knowledge management, standards, Systems Biology

Dynamical modelling of T cell co-inhibitory pathways to predict anti-tumour responses to checkpoint inhibitors

Celine Hernandez [*†] [1], Aurélien Naldi [1], Wassim Abou-Jaoudé [1],
Guillaume Voisinne [2], Romain Roncagalli [2], Bernard Malissen [2,3],
Morgane Thomas-Chollier [1], Denis Thieffry[‡] [1]

[1] Institut de biologie de l'école normale supérieure (IBENS) – École normale supérieure [ENS] - Paris, Inserm : U1024, CNRS : UMR8197 – 46, rue d'Ulm, 75005 Paris, France

[2] Centre dÍmmunologie de Marseille - Luminy (CIML) – Aix Marseille Université : UM2, Institut National de la Santé et de la Recherche Médicale :
UMR$_{S}$1104, $CentreNationaldelaRechercheScientifique$:
$UMR7280 - -ParcscientifiqueettechnologiquedeLuminy - 163, avenuedeLuminy - Case906 -$
$13288Marseillecedex09, France$

[3] Centre dÍmmunophénomique (CIPHE) – Aix Marseille Université : US012, Institut National de la Santé et de la Recherche Médicale : US012, Centre National de la Recherche Scientifique : UMS3367 – Parc Scientifique et Technologique de Luminy - 163 avenue de Luminy - Case 936 - 13288 Marseille Cedex 9, France

In recent years, it has been recognised that T cells often display a reduced ability to eliminate cancer cells and that expression of co-inhibitors at their surface accounts for their compromised function. By blocking the functions of these co-inhibitors, therapeutic antibodies (checkpoint inhibitors) have become standard treatment for metastatic melanoma [1], leading to a revival in the study of T cell co-inhibitors. However, our understanding of the immunobiology of T cell co-inhibitors and of their harmful role during anti-tumour responses remains fragmentary. Despite a few biochemical studies, a mechanistic understanding at the system-level of the modulation of T cell function by co-inhibitors has remained elusive.
To overcome these limitations, we aim at delineating the mechanisms through which co-inhibitory molecules such as PD-1 and CTLA-4 impede T cell functions at the system-level. To reach our goal, we combine high-throughput analysis with computational methods to map TCR co-signalling pathways and predict cell responses to perturbations.

First, we focused on the development of comprehensive annotated molecular maps based on the curation of scientific literature, in parallel with automated queries to public databases and protein-protein graph reconstruction. Next, using the software GINsim [2], these maps and protein networks were translated into a regulatory graph integrating current knowledge. The challenge is then to properly model concurrent intracellular processes, along with feedback control mechanisms. To cope with this complexity, we first modelled network modules using a Rule-based formalism [3], in order to explore concurrent biological hypotheses and specify log-

[*] Speaker

[†] Corresponding author: celine.hernandez@ens.fr

[‡] Corresponding author: thieffry@ens.fr

ical rules recapitulating observed component behaviour. These modules will be integrated into a single logical model and used to predict cell response to single or multiple perturbations, and thereby pave the way to the delineation of novel experiments, which will in turn be used to refine the maps and model.

This integrated system-level view of the mechanisms of action of key T cell co-inhibitors in cancer will further provide a rationale for designing and evaluating drugs targeting T cell co-inhibitory pathways in anti-cancer immunotherapy.

References

1. Simpson TR, Li F, Montalvo-Ortiz W, Sepulveda MA, Bergerhoff K, Arce F, Roddie C, Henry JY, Yagita H, Wolchok JD, Peggs KS, Ravetch JV, Allison JP, Quezada SA (2013). Fc-dependent depletion of tumor-infiltrating regulatory T cells co-defines the efficacy of anti-CTLA-4 therapy against melanoma. *The Journal of experimental medicine* **210**(9): 1695–710.

2. http://www.ginsim.org

3. Feret J, Danos V, Krivine J, Harmer R, Fontana W (2009). Internal coarse-graining of molecular systems. *Proceedings of the National Academy of Sciences of the USA* **106**(16): 6453-8.

Keywords: immunology, checkpoint inhibitor, PD1, CTLA4, modelling, logical modelling, dynamical modelling

The vanishing Gram-positives: a third lineage with outer membranes in the Firmicutes

Najwa Taib * [1,2], Daniel Poppleton [1], Guillaume Borrel [1], Christophe Beloin [3], Simonetta Gribaldo† [1]

[1] Biologie Evolutive de la Cellule Microbienne, Département de Microbiologie - Institut Pasteur – Institut Pasteur de Paris – France

[2] HUB Bioinformatique et Biostatistique, C3BI - Institut Pasteur – Institut Pasteur de Paris – France

[3] Génétique des biofilms, Département de Microbiologie - Institut Pasteur – Institut Pasteur de Paris – France

The bacterial cell envelope is one of the most ancient features of life; yet, most aspects of its evolutionary history remain obscure. Because the phylogenetic relationships among monoderm and diderm bacterial phyla are ill-resolved, the details of such transition have been elusive. In this respect, the existence of both monoderm and diderm members within the same bacterial phylum represents a unique opportunity to clarify this issue. This is the case of the Firmicutes, which include two clades that display an outer membrane, the Negativicutes and the Halanaerobiales.

We have recently put forward the hypothesis that the diderm envelope of Negativicutes and Halanaerobiales is an ancestral characteristic of the Firmicutes that was retained only in these two lineages, while it was lost multiple times independently during the diversification of this phylum to give rise to the classical monoderm cell envelope architecture (Antunes et al. 2016). This hypothesis opens the possibility that additional diderm lineages may be present in the Firmicutes. Indeed, the first member of another diderm Firmicute lineage, the Limnochordales, was isolated from a brackish meromictic lake (Watanabe *et al.*, 2015). The genome of *Limnochorda pilosa* revealed the presence of classical OM markers, consistent with an ultrastructural evidence for a diderm cell envelope (Watanabe *et al.*, 2016).

Here, we searched to enrich the genomic data for the Limnochordales and the other two diderm lineages by taking advantage of the recent release of nearly 1500 new genomes from uncultured Firmicutes that were assembled from available metagenomics databanks (Parks et al., 2017). First, we searched for conserved taxonomic markers which were included in an updated reference phylogeny of the Firmicutes. We identified 1 new Halanaerobiale, 40 new Negativicutes, and 40 new Limnochordales. This helped robustly placing the Limnochordales as a third independent and deep-branching diderm lineage. All Limnochordales possess a large cluster of OM markers previously identified in Halanaerobiales and Negativicutes, indicating that it is a conserved feature of diderm Firmicutes and providing key information for experimental functional validations. Moreover, an updated phylogeny of four conserved LPS genes strongly supports the hypothesis of a diderm ancestor for the Firmicutes.

*Speaker

†Corresponding author: simonetta.gribaldo@pasteur.fr

Our results show that the presence of an OM in the Firmicutes is more widespread than previously thought, making this phylum a truly mixture of diderms and monoderms and opening the way for further analysis of this important cellular transition in bacterial evolution.

Keywords: Cell envelope, Outer membrane, Firmicutes, Evolution

Characterization of mutations that dysregulate driver microRNAs in breast cancer

Jaime Castro-Mondragon [1], Miriam Aure [2], Evita Lindholm [2], Ole Christian Lingjaerde [2,3], Anita Langerod [2], Anne-Lise Borresen-Dale [2], Vessela Kristensen [2], Anthony Mathelier *† [2,4]

[1] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway (NCMM) – Norway

[2] Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway – Norway

[3] Department of Computer Science, University of Oslo, 0316 Oslo, Norway – Norway

[4] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway – Norway

MicroRNAs (miRNAs) represent a class of small (_˜21nt) noncoding RNAs that regulate post-transcriptional gene expression through mRNA degradation or translational repression. Their biogenesis comprises three main steps, (i) the transcription of a primary miRNA (pri-miRNA), which can be of hundreds of base pairs to mega-base pairs long, (ii) the cleavage of a miRNA precursor (pre-miRNA) of about 70 bp by Drosha, and (iii) the cleavage of the pre-miRNA by Dicer, which will produce the mature miRNA. A single miRNA may target tens (or even hundreds) of mRNAs to regulate key cellular processes such as differentiation, growth, and apoptosis. Hence, miRNA expression must be accurately controlled and alteration of their regulation has already been linked to diseases such as cancer. Despite active research on the identification of miRNAs and their targets, the understanding of their transcriptional regulation has been limited by a lack of knowledge regarding the location of their transcription start sites (TSSs) and associated cis-regulatory sequences (e.g. promoters and enhancers). Recent studies have independently predicted human miRNA-associated TSSs genome-wide across _˜300 samples from different cell types. These data provide an unprecedented opportunity to analyze miRNA transcriptional regulation and its alteration in cancer.

With a wealth of individual molecular data available from cancer patients obtained by international consortia (e.g., BASIS, ICGC, METABRIC, TCGA), it is critical to integrate multiple layers of information to study the impact of mutations on the dysregulation of the regulatory program in cancer cells. As cancer is a disease of dysregulation, our project aims at detecting mutations in pri-miRNA and their *cis*-regulatory elements that dysregulate miRNA expression (*cis* effect of mutations) with a cascading effect on the dysregulation of their target mRNAs (*trans* effect of mutations).

We applied a previously developed probabilistic framework, xseq, to relate specific mutations to expression disruption (up- or down-regulation) of miRNAs and their targets on data from

*Speaker

†Corresponding author: anthony.mathelier@ncmm.uio.no

294 breast cancer patients from the BASIS consortium. Specifically, our datasets comprised trios of (i) somatic mutations extracted from whole genome sequencing (WGS) of normal and tumour samples, (ii) RNA-seq, and (iii) miRNA microarrays from tumours for each patient. When considering somatic point mutations and small ($<$ 200 nt) insertion/deletion lying within pri-miRNAs, xseq identified 12 miRNAs as down-regulated in the patients harbouring somatic mutations. Five out of the twelve selected miRNAs are found in an imprinted loci (chr 14q32) containing the largest cluster of human miRNAs ($>$ 50), which have been previously demonstrated to be dysregulated in several cancer types, including breast cancer. Using an independent cohort (METABRIC; 1282 patients), we observed that down-regulation of the 12 miRNAs is associated with worse prognosis.

As clusters of enhancers, also known as super-enhancers, have been shown to be involved in miRNA processing, we independently considered somatic mutations lying within these clusters and assessed their likely assocations with altered expression of the miRNAs they regulate. This analysis highlighted two miRNAs that have previously been associated with breast cancer but for which we now provide some insights on the molecular mechanisms underlying their dysregulation.

We will extend the xseq tool to consider the *trans* effect of the altered expression of these miRNAs in breast cancer patients. Specifically, the approach will assess the likely association between the presence of mutations within a pri-miRNA or associated regulatory elements with observed deviations from neutral expression of miRNA and target mRNA expression. The analysis in *trans* will allow us to integrate information of both transcriptional and post-transcriptional regulation of gene expression to further shed light into the molecular mechanisms involved in carcinogenesis.

Keywords: microRNA, somatic mutations, cancer, systems biology

The CoLoMoTo Interactive Notebook: Accessible and Reproducible Computational Analyses for Qualitative Biological Networks

Aurélien Naldi [1], Celine Hernandez [1], Nicolas Levy , Gautier Stoll , Pedro Monteiro , Claudine Chaouiya , Tomáš Helikar , Andrei Zinovyev [2], Laurence Calzone , Sarah Cohen-Boulakia [3], Denis Thieffry [1], Loïc Paulevé * [4]

[1] Institut de biologie de l'école normale supérieure (IBENS) – École normale supérieure [ENS] - Paris, Inserm : U1024, CNRS : UMR8197 – 46, rue d'Ulm, 75005 Paris, France

[2] Cancer et génôme: Bioinformatique, biostatistiques et épidémiologie d'un système complexe – Inserm : U900, Institut Curie, MINES ParisTech - École nationale supérieure des mines de Paris – 26 rue d'Ulm - 75248 Paris cedex 05, France

[3] Laboratoire de Recherche en Informatique (LRI) – Université Paris-Sud - Paris 11, Centre National de la Recherche Scientifique : UMR8623 – LRI - Bâtiments 650-660 Université Paris-Sud 91405 Orsay Cedex, France

[4] Laboratoire de Recherche en Informatique (LRI) – CNRS : UMR8623, Université Paris Sud – LRI - Bâtiments 650-660 Université Paris-Sud 91405 Orsay Cedex, France

Analysing models of biological networks typically relies on workflows in which different software tools with sensitive parameters are chained together, many times with additional manual steps. The accessibility and reproducibility of such workflows is challenging, as publications often overlook analysis details, and because some of these tools may be difficult to install, and/or have a steep learning curve.

The CoLoMoTo Interactive Notebook provides a unified environment to edit, execute, share, and reproduce analyses of qualitative models of biological networks. This framework combines the power of different technologies to ensure repeatability and to reduce users' learning curve of these technologies. The framework is distributed as a Docker image with the tools ready to be run without any installation step besides Docker, and is available on Linux, macOS, and Microsoft Windows. The embedded computational workflows are edited with through a Jupyter web interface, enabling the inclusion of textual annotations, along with the explicit code to execute, as well as the visualisation of the results. The resulting notebook files can then be shared and re-executed in the same environment. To date, the CoLoMoTo Interactive Notebook provides access to software tools including GINsim, BioLQM, Pint, MaBoSS, and Cell Collective for the modelling and analysis of Boolean and multi-valued networks. More tools will be included in the future. We developed a Python interface for each of these tools to offer a unified and seamless integration in the Jupyter web interface and ease the chaining of complementary analyses.

*Speaker

Keywords: computational systems biology, reproducibility, model analysis, Boolean networks, Python programming language

In-depth analysis of the impact of transposable elements on genome assembly quality

Rémy Costa *† [1], Yasmine Mansour‡ [2,3], Annie Chateau§ [3,4], Anna-Sophie Fiston-Lavier¶ [2]

[1] Institut des Sciences de lÉvolution de Montpellier (ISEM) – Université de Montpellier, Institut de recherche pour le développement [IRD] : UR226, Centre National de la Recherche Scientifique : UMR5554 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France

[2] Institut des Sciences de l'Evolution - Montpellier (ISEM) – CNRS : UMR5554, Institut de recherche pour le développement [IRD] : UMR226 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France

[3] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université Montpellier II - Sciences et techniques, CNRS : UMR5506 – CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France

[4] Institut de Biologie Computationnelle (IBC) – CNRS : UMR5506, Université Montpellier II - Sciences et techniques – 95 rue de la Galéra, 34095 Montpellier, France

Genome assembly has become crucial for conducting genomic studies in various field as environment, health, genetics, evolution and many more. Recent studies highlighted the impact of assembly quality on result interpretations [1]. While efficiency of bioinformatic tools used for assembly is increasing, errors of sequence construction from contigous short reads persist. One of the known sources of errors is repeated elements.
The presence of repeated elements can induce (i) chimeric contigs due to collapsed repeats and (ii) assembly breaks. Among repeated elements, transposable elements (TEs) are ubiquitous sequences, *i.e.* detected in the vast majority of sequenced genomes, and make up for a large fraction of them (*e.g.* up to 90% for the maize genome) [2]. A variety of TEs can be identified. They are classified according to their transposition mecanisms and sequence properties [3].

The recently sequenced and assembled genome of *Ambystoma mexicanum* (Mexican axolotl) shows that up to 97% of contigs encompass TEs at their ends. Analysis of these TEs showed that they are recent (sharing a high sequence identity) and abundant (present in numerous copies). Such active TEs mainly correspond to a specific group : LTR retrotransposons [4]. Even if advanced sequencing technologies has improved assembly quality such as long read sequencing, no short read based approaches allow investigating in-depth analysis of disruptive TEs. We expect TE-rich genomes to be harder to assemble, and specific type of TEs to cause more errors than others. Recent and long TEs with a high copy number should induce more assembly biases. As TEs do not insert homogenously in the genome, we also expect regions

*Speaker

†Corresponding author: remy.costa@etu.umontpellier.fr

‡Corresponding author: yasmine.mansour@umontpellier.fr

§Corresponding author: annie.chateau@lirmm.fr

¶Corresponding author: anna-sophie.fiston-lavier@umontpellier.fr enriched in TEs to be more challenging to assemble.

Here we aim to test our hypotheses by estimating the impact of TEs on assembly quality through identifying the most disruptive TE types and analyzing the impact of TE density on the assembly quality. For that, we will use an approach based on assembly simulation by controlling TE features in the *Drosophila melanogaster* genome. This genome harbors one of the highest quality genomic sequences and annotations. Our results should help improving the process of genome assembly by taking advantage of the TE information.

Mahul Chakraborty, Nicholas W. Vankuren, Roy Zhao, Xinwen Zhang, Shannon Kalsow, and J. J. Emerson. Hidden genetic variation shapes the structure of functional elements in Drosophila. Nature Genetics, 50(1) :20–25, 2018.

Dario Copetti and Rod A. Wing. The Dark Side of the Genome : Revealing the Native Transposable Element/Repeat Content of Eukaryotic Genomes. Molecular Plant, 9(12) :1664–1666, 2016.

Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel, and Alan H. Schulman. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics, 8(12) :973–982, 2007.

Sergej Nowoshilow, Siegfried Schloissnig, Ji Feng Fei, Andreas Dahl, Andy W.C. Pang, Martin Pippel, Sylke Winkler, Alex R. Hastie, George Young, Juliana G. Roscito, Francisco Falcon, Dunja Knapp, Sean Powell, Alfredo Cruz, Han Cao, Bianca Habermann, Michael Hiller, Elly M. Tanaka, and Eugene W. Myers. The axolotl genome and the evolution of key tissue formation regulators. Nature, 554(7690) :50–55, 2018.

Keywords: transposable elements, genome assembly, high, throughput sequencing, NGS

A logical approach to identify Boolean Networks that model cell differentiation

Stéphanie Chevalier *† 1, Andrei Zinovyev 2, Christine Froidevaux 1, Loïc Paulevé‡ 3

1 Laboratoire de Recherche en Informatique (LRI) – Université Paris-Sud - Paris 11, Institut National de Recherche en Informatique et en Automatique, Centre National de la Recherche Scientifique : UMR8623, CentraleSupélec – LRI - Bâtiments 650-660 Université Paris-Sud 91405 Orsay Cedex, France
2 Cancer et génôme: Bioinformatique, biostatistiques et épidémiologie d'un système complexe – Inserm : U900, Institut Curie, MINES ParisTech - École nationale supérieure des mines de Paris – 26 rue d'Ulm - 75248 Paris cedex 05, France
3 Laboratoire de Recherche en Informatique (LRI) – CNRS : UMR8623, Université Paris Sud – LRI - Bâtiments 650-660 Université Paris-Sud 91405 Orsay Cedex, France

Context

The gene interactions can be represented by networks, where gene activation and inhibition are modeled. By predicting the result of existing perturbations in diseases, these computational models of gene regulation bring a promising framework to suggest new therapeutic targets. However, the formal definition of causal networks that model the growth of a studied cell line is a major challenge.

Among the several modeling paradigms for networks, this work focus on qualitative models of dynamics, especially Boolean networks, where the activity of genes is seen as simply ON or OFF. Boolean networks allow capturing essential dynamical features, such as steady states and differentiation processes, whereas requiring few parameters compared to quantitative models. Moreover, by being more abstract, Boolean networks allow both to address large-scale networks and to derive robust predictions, which depend very little on precise quantitative features.

The inference of Boolean networks has been addressed in the literature for steady-state data and time-series data, on cell lines subjects to perturbations. However, no method allows to take into account differentiation features. Hence we explore model inference for cell differentiation to provide a scalable framework for the systematic identification of differentiation models. Considering the nature of differentiation data influences the model checking, the first step was to formalize the various contexts of cell differentiation data, according to the stability of the measured cells and the measurements representativeness.

The systematic approach aims at reducing biases of over-fitting, by understanding which part of the networks are crucial to reproduce the data (should be common to all found models), and part of the networks which are non-identifiable given the data (high variability among the admissible models). This approach rules out any try-and-test approaches which would simply iterate over

*Speaker

†Corresponding author: stephanie.chevalier@u-psud.fr

‡Corresponding author: loic.pauleve@lri.fr

all possible models and independently test their validity. Instead, we rely on logic programming to express in a same abstract model both the possible networks and the constraints they have to satisfy to fit the data depending on the contexts previously defined. The logical characterization of the network inference starts as an extension of a prior work [2] on model inference from time series data using Answer-Set Programming (ASP) that enumerates the solutions of the logical program.

After a brief description of the Boolean networks, we introduce different levels of interpretation of the differentiation data and we describe ensuing formal constraints. We apply our method on a network that models the central nervous system differentiation.

Boolean Networks

A Boolean Network (BN) is a pair of two sets [4] : a finite set of components (genes), and the corresponding set of Boolean functions.

A component at time t takes the value either 1 (expressed) or 0 (not expressed). A state contains the overall expression level of all components of the Boolean network at time step t. The state of each gene can be updated synchronously or asynchronously during a transition, respecting the Boolean functions. In this work we consider the asynchronous Boolean network (ABN) : one component is updated at a time.

A consecutive sequence of states obtained by state transitions is called a trajectory. A reachable state from a state x of the BN is a state that belong to a trajectory from x. State transition in ABN is non-deterministic : from a state, it is possible that there are several directly reachable states. Ultimately a trajectory reaches an attractor, which is defined as follows : the set of reachable states from a state of the attractor is the set of states of the attractor itself. The attractor is called a fixed point if it is a single state, and cyclic attractor otherwise. In our context, the attractors correspond to the long-term behavior of the cells and thus represent the phenotypes. A basin of attraction of an attractor is the set of all states that reach this attractor.

Various contexts of cell differentiation data

The differentiation data are gene expression measurements on cell populations or single cells, that reach various phenotypes depending on activation or inhibition of genes. To interpret it, we have formally defined what differentiation is, and we have observed that it corresponds to different experimental contexts, which influence the constraints to identify the Boolean networks that reproduce the data :

1) The first level of contexts refers to hypothesis on the cell stability : it distinguishes whether measured cells were in a stable state (”steady observation”) or not (”non-steady observation”).
2) The second level of contexts refers to hypothesis on the measurement representativeness : it distinguishes whether the sample is representative of the whole diversity of the population (”universal observation”) or not (”existential observation”).
3) The third level of contexts refers to a particularity of the steady data : the measurement can correspond to a single state (”fixed point”) or a cycle of states (”cyclic attractor”). And in this last specific case, another level of hypothesis can be made : the markers measured can be permanent or periodic in the attractor’s states.

Depending on the context, the interpretation in constraints for the Boolean networks is different.

From interpretation to formal constraints

To infer qualitative models of differentiation mathematically expressed as Boolean networks, the first strategy would be to check the reachable states from the differentiated states. But checking the reachability of a state from another is a standard model-checking task, known to have a limited scalability due to its theoretical complexity (PSPACE-complete[1]) [2]. Hence our first concern is to translate data into constraints that have to be satisfied by the dynamics of a Boolean network model in order to be considered as a valid candidate model. Constraints depend on the contexts of the differentiation, and have to be expressed using logical programming with the aim of filter out quickly some true negatives (thanks to necessary conditions) and true positives (thanks to sufficient conditions), leaving the computational demanding validation to a subset of putative models.

Our current work focuses on the implementation of sufficient conditions. For instance let us consider a sufficient condition in the non-steady context, which consists of an over-approximation of the reachable states : if the over-approximation from the states that are differentiated is disjoined, we can already confirm that the Boolean network reproduces the differentiation, without a costlier check. The strategy actually slightly differs according to the hypothesis "universal/existential data", but the main current challenge is to translate it into logic programming, to extend the tool Caspots[2] that already has constraints on reachability.

The context of steady cells means in the Boolean network context that the observed states are in attractors. For the case of fixed point, an option has been added to Caspots, that allows to indicate time points as fixed points. For the case of cyclic attractor, constraint can be expressed in temporal logic (CTL) that allows a quite efficient network check.

Applications

To train our method we produce simulated data adapted to the manually-curated minimal network for central nervous system differentiation [3]. This network mainly consists of two steps of two mutually-inhibited fate decision genes : the first step determines the specialization in neuron or glia where the second step determines that of astrocyte and oligodendrocyte.

Infer on this 5-node network by considering time series data highlights 256 compatible Boolean networks. When we specify the specialization in neuron, glia and astrocyte as 3 fixed points thanks to our new option integrated in caspots, this constraint brings out that only 4 of these Boolean networks are actually compatible with the data.

Thus, when it is known that time points of time series data correspond to steady states, the fixed point constraint is relevant to discriminate efficiently Boolean networks.

Perspectives

The short-term focus is to find other constraints to implement in ASP in the differentiation contexts described, so as to infer network efficiently, even under realistic conditions with large gene networks. But other axes complete this work.

A first topic is that, up to now, we have only considered the existence of differentiation processes. We must also cover the experimental situation where differentiation leads to a change in the probabilities of reaching phenotypes.

Then in a longer-term view, being able to identify networks that reproduce the observed differentiation on time series data is a first step in the goal of network inference about CRISPR-Cas9 mutation data, for which the only dynamic information could be the phenotype of the cell population (proliferation, apoptosis, ...).

References

Allan Cheng, Javier Esparza, and Jens Palsberg. Complexity results for 1-safe nets. Theoretical Computer Science, 147(1) :117 – 136, 1995.

Max Ostrowski, Lo´ic Paulevé, Torsten Schaub, Anne Siegel, and Carito Guziolowski. Boolean Network Identification from Perturbation Time Series Data combining Dynamics Abstraction and Logic Programming. BioSystems, 2016.

Xiaojie Qiu, Shanshan Ding, and Tieliu Shi. From understanding the development landscape of the canonical fate-switch pair to constructing a dynamic landscape for two-step neural differentiation. PLOS ONE, 7(12) :1–14, 12 2012.

René Thomas. Boolean formalization of genetic control circuits. Journal of Theoretical Biology, 42(3) :563–585, 1973.

Keywords: boolean network, cell differentiation, network inference

Alienness vs. Predictor (AvP): fast and robust detection of horizontal gene transfers across the tree of life

Corinne Rancurel *† [1], Solène Granjeon-Noriot [1], Etienne G J. Danchin * ‡ [1], Georgios D. Koutsovoulos§ [1]

[1] Institut Sophia Agrobiotech [Sophia Antipolis] (ISA) – Institut National de la Recherche Agronomique : UMR1355, Université Nice Sophia Antipolis : UMR7254, Centre National de la Recherche Scientifique : UMR7254 – INRA Centre de recherche Provence-Alpes-Côte dÁzur 400, route des Chappes BP 167 06903 Sophia Antipolis Cedex, France

Horizontal gene transfer (HGT) is a process by which an organism integrates coding genetic material from another organism by means other than "vertical" inheritance from an ancestor. HGT has long been recognized as an important mechanism shaping both the genome and biology of prokaryotes (e.g. acquisition of antibiotic resistance). Although the impact of HGT in eukaryotes is probably lower, recent evidences suggest this mechanism can bring novelty in eukaryotic biology too (e.g. adaptation to parasitism) (Danchin et al. 2016). Furthermore, substantial proportions of the gene sets of several eukaryotic species (e.g. fungi or rotifers) have likely been acquired via HGT. These observations combined with the availability of more and more whole genomes across the tree of life have raised interest in methods able to rapidly and robustly detect HGT in large datasets.
Here, we present Alienness vs. Predictor (AvP), a suite of bioinformatics tools that allow rapid and reliable detection of HGT in any genome from any potential donor. AvP is made of two main components: (i) Alienness that allows to rapidly detect putative HGT at high-throughput, based on BLAST searches and (ii) Predictor that performs automatic phylogenies and searches for tree topologies supporting HGT from Alienness candidates. AvP can be used as a suite combining the two components, but each component can also be used separately depending on the needs of the user.

The two components are described in more details below.

Alienness (Rancurel et al. 2017) is a taxonomy-aware web application that parses BLAST results against public libraries to rapidly identify candidate HGT in a genome of interest. Alienness takes as input the result of a BLAST of a whole proteome of interest against any NCBI protein library. The user defines recipient (e.g. metazoan) and donor (e.g. bacteria, fungi) branches of interest in the NCBI taxonomy. Based on the best blast E-values of candidate donor and recipient taxa, Alienness calculates an Alien Index (AI) for each query protein. Our method uses the Alien Index metrics as described in (Gladyshev, et al, 2008) to detect a significant gap

*Speaker

†Corresponding author: corinne.rancurel@inra.fr

‡Corresponding author: etienne.danchin@inra.fr

§Corresponding author: georgios.koutsovoulos@inra.fr of magnitude in E-values between candidate donor and recipient taxa. An AI > 0 indicates a better hit to candidate donor than recipient taxa and a possible HGT. Higher AI represents a higher gap of E-values between candidate donor and recipient and a more likely HGT. The Alienness tool presents a fast and interesting method to rapidly identify candidate HGT and narrow down the number of genes to be analyzed afterwards.

The gold standard methodology to verify the validity of a candidate horizontal transfer is phylogenetic analysis. The principle is to build a phylogenetic trees of the genes and compare the obtained topology to the expected reference species tree. By identifying inconsistencies between the gene and species trees, it is possible to identify topologies supporting an HGT event. This is exactly what the Predictor component of AvP performs. By default, Predictor first groups genes that are duplicated or form a multigene family into clusters (based on similarity of their list of BLAST hits). Then, for each cluster with at least one gene returning an AI> 0, Predictor performs a multiple sequence alignment followed by Maximum-Likelihood phylogeny. The first step consists in retrieving a non-redundant pool of BLAST hits from the whole cluster. Then the pool of proteins is aligned with MAFFT (Katoh & Stanley. 2013) with an optional alignment trimming with TrimAl (Capella-Gutiérrez et al. 2009). Predictor then proposes three options to perform Maximum Likelihood phylogenetic analyses, depending on the need for speed vs. accuracy: FastTree (Price et al. 2010), IQ-Tree (Nguyen et al. 2015) and raxml-ng (https://github.com/amkozlov/raxml-ng). By default, the LG+G model of evolution is selected for FastTree and raxml-ng, while IQ-Tree performs a model selection from WAG, LG, and JTT testing E, I, G, and R rates. However, the user can select any other model supported by the respective program. Because the reconstructed phylogenetic tree is annotated for taxonomy using a tagging system, it is possible to decipher the taxonomic identity of the sequences falling in the same branch than the protein of interest. Basically, if the protein studied falls in a group exclusively constituted by proteins that belong to donor taxonomic groups, it is likely that the gene comes from a horizontal gene transfer. If the proteins falls in a group containing only proteins from sister taxa, then the tree does not support an hypothesis of HGT. Other cases, where the protein of interest falls in a group that contain both sister taxa and donor taxa are deemed complex cases and tagged as such. In any case, the user can visualize the produced annotated trees to make a more informed decision.

HGT vs. Contamination

Obviously, contamination can easily be misinterpreted as an HGT because it will return a high AI score and also provide a tree topology indicative of HGT (contaminated gene among a clade composed of the donor species). It is thus important to distinguish possible contamination from likely HGT (Koutsovoulos et al. 2016). AvP deals with this problem by providing several alerts and indicators for possible contamination. First, in the Alienness step, all the genes that return at least 70% identity to candidate donors are tagged as possible contamination in the rest of the analysis. Second, in Predictor, if a GFF for the query genome under consideration is provided, the neighborhood of the gene under consideration is analyzed. If bona fide genes from the query species (typically with an AI< 0) are present in the vicinity of the gene under consideration, then contamination is less likely. In the opposite, if the vicinity of the gene is exclusively composed of other genes with an AI> 0 then contamination is the more likely explanation. Furthermore, Predictor also provides information about the presence of spliceosomal introns in the gene of interest. In the particular case of HGT of prokaryotic origin in eukaryotes, this information is useful to support HGT rather than contamination. Finally, expression data, if available, can also be incorporated within AvP to bring functional support to the HGT event.

Future directions:

Even if phylogenetic analysis is considered the gold standard to support the HGT hypothesis, it may also suffer from confounding effects such as multiple losses of an ancestral gene, reciprocal loss of out-paralogs or incomplete lineage sorting. To address these points we plan to develop an additional layer of analysis : hypothesis testing. We will construct alternative constrained topology (e.g. requiring separate monophyly of donor and receptor groups) and compare the likelihood of the constrained topology to that of the topology supporting the HGT hypothesis using CONSEL (Shimodaira and Hasegawa. 2001) or similar metrics.

Availability:

The Alienness component of AvP is already publicly available for free at the following URL: alienness.sophia.inra.fr

The Predictor component is still under development and test in our lab and will be made publicly available as a downloadable package in a first instance, then deployed on a publicly-available server in the future.

Bibliography:

Danchin E.G.J. Lateral gene transfer in eukaryotes: tip of the iceberg or of the ice cube? BMC Biology. 2016

Rancurel, C., Legrand, L., Danchin, E.G.J. Alienness: Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life. Genes 2017

Gladyshev, E. A., Meselson M., Arkhipova I.R. Massive Horizontal Gene Transfer in Bdelloid Rotifers. Science 2008

Koutsovoulos G., Kumar S., Laetsch D.R., Stevens L., Daub J.,Conlon C., Maroon H., Thomas F., Aboobaker A.A., Blaxter M. No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. PNAS 2016

Katoh, K., et D. M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 2013

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, et Toni Gabaldón. TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. Bioinformatics. 2009

Price, Morgan N., Paramvir S. Dehal, et Adam P. Arkin. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE. 2010

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, et Bui Quang Minh. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2015
Shimodaira, H., et M. Hasegawa. CONSEL: For Assessing the Confidence of Phylogenetic Tree Selection. Bioinformatics. 2001

Keywords: horizontal gene transfer, lateral gene transfer, HGT, alien index, alienness, AVP, phylogeny

bioLQM: a java library for the manipulation and conversion of logical qualitative models of biological networks

Aurélien Naldi * [1]

[1] Institut de Biologie de l'ENS (IBENS) – Institut National de la Santé et de la Recherche Médicale - INSERM : U1024, Centre National de la Recherche Scientifique - CNRS : UMR8197 – 46 rue d'Ulm, 75005 Paris, France

Qualitative dynamical models have been increasingly used for the analysis of complex cellular networks, leading to the development of a number of complementary software tools. The CoLoMoTo consortium gathers many groups involved in the development of software modelling tools tools, as well as of various kinds of biological applications [1]. The CoLoMoTo consortium has recently defined the SBML qual format [2] to ease model exchange and hence take advantage of complementary features of the various tools.
The bioLQM toolkit [3] provides conversion bridges between SBML qual and several other model definition formats, facilitating the design of workflows combining multiple tools. An extensible architecture facilitates the addition of new model definition formats.

The SBML qual format supports the definition of multi-valued models, however some analysis tools are limited to Boolean models only. To bypass this limitation, bioLQM supports the transformation of multi-valued models into Boolean models with an identical dynamical behaviour by mapping multi-valued components on Boolean ones.

Before the conversion of a model, bioLQM supports the optional definition of one or several model transformations, such as the booleanization step mentioned above. Model perturbations (also called mutations) are another popular model transformation, enforcing the activity of a specific component or interaction. BioLQM also supports several model reduction methods to ease the analysis of large models. Finally, bioLQM provides a framework for the development of novel analysis tools. The current version implements the usual updating modes for model simulation (notably synchronous, asynchronous, and random asynchronous), as well as some static analysis features for the identification of attractors.

The bioLQM software can be integrated into analysis workflows through command line and scripting interfaces. As a Java library, it further provides core data structures to the GINsim and EpiLog interactive tools, which supply graphical interfaces and additional analysis methods for cellular and multi-cellular qualitative models.

References

*Speaker

1. http://www.colomoto.org/

2. http://sbml.org/Documents/Specifications/SBML_Level_3/Packages/qual
3. https://github.com/colomoto/biolqm

Keywords: Logical modeling, regulatory networks

Modelling gene regulatory networks in oncogene-induced senescence

José Américo Nabuco Leva Ferreira De Freitas * [1], Pierre-François Roux [1], Ricardo Iván Martínez-Zamudio [1], Lucas Robinson [1], Gregory Doré [1], Nir Rozenblum [1], Benno Schwikowski [2], Oliver Bischof [1]

[1] Laboratory of Nuclear Organization and Oncogenesis – Institut Pasteur de Paris – 28, rue du Dr. Roux F-75015 Paris, France

[2] Systems Biology Laboratory – Institut Pasteur de Paris – 25, rue du Dr. Roux F-75015 Paris, France

Cellular senescence (CS) is a cell fate that arrests cell proliferation in response to numerous stresses most notably oncogenes such as RAS [1]. The senescence arrest is essentially permanent and accompanied by widespread changes in chromatin structure, metabolism and gene expression, including a senescence-associated secretory phenotype (SASP) – the expression and secretion of inflammatory cytokines, growth factors, proteases and other molecules that can promote senescence clearance, reinforce senescent phenotypes (intracrine-autocrine senescence) or alter tissue microenvironments (paracrine senescence). An emerging paradigm stipulates that senescent cells are major contributors to health and age-related illnesses, particularly cancer, by virtue of the SASP. As such, research on therapeutic strategies exploiting senescence targeting to improve healthspan has gained enormous momentum in recent years [2]. Recent studies show that CS onset is highly dynamic, with several markers evolving differently across time and tissue type [3]. Thus, a firm understanding of the time-dependent interactions of CS regulators is essential to infer their hierarchical order to observe causality in their dynamics and provide the means to design predictive models.

The phenotypic and transcriptomic changes that occur during CS can be interpreted as transitions in a high-dimensional state space, where each state component is equivalent to molecular species concentration or gene expression level, as shown in cellular differentiation [4] and cancer development [5]. This high-dimensional state space is constituted by diverse attractors, that correspond to stable cell phenotypes and are maintained by the activity of regulatory negative feedback loops [6]. Attractors can be destabilized by the action of external stimuli, inducing the displacement of the system to a newly formed attractor, which corresponds to the new acquired phenotype and its associated gene expression levels.

High-dimensional landscape changes are associated to fluctuations in the interactions between genes, i.e., the topology of the underlying gene-regulatory network (GRN) [4]. For instance, densely connected GRNs are associated to attractors, while modular networks correspond to smoother landscapes [7]. To model the transcriptomic landscape transitions occurring during the establishment of CS, we applied the implicit sparse identification of nonlinear dynamics algorithm (implicit-SINDy) [8] to a time-course gene expression data on fibroblasts undergoing RAS-induced senescence. The implicit-SINDy algorithm allow us to determine both model structure and parameters, i.e., the GRN interactions and their respective intensities. Briefly,

[*] Speaker

the algorithm consists of three steps: creating a matrix Θ whose columns consist of a library of arbitrary nonlinear functions over the time-course data; computing the matrix null space, which defines the possible sets of coefficients that fit the nonlinear libraries to the system dynamics; and finding the sparsest vector in this subspace, resulting in a parsimonious model that fits our experimental results. The system's implicit formulation enables our model to follow rational functions, such as the Michaelis-Menten kinetics equation, known to describe several enzymatic reactions. We also address the high dimensionality of transcriptomic data, i.e. the "large p, small n" problem, by generating the matrix Θ over the temporal modes obtained by the proper orthogonal decomposition (POD) [10].

Recent advances in OMICS technologies allowed for the acquisition of comprehensive genome-wide datasets to chart the (epi)genomic, transcriptomic, metabolomic and proteomic landscape of cells, tissues, and organs. These datasets revealed the immense complexity of cellular regulatory processes, that are regulated by intricate gene-regulatory networks (GRN) rather than by individual genes. In order to describe and predict GRN activity, we need systematic and integrative approaches that also consider the temporal evolution of its constituents. Our proposed predictive modeling approach will provide a deeper understanding of cellular senescence and has the potential to lay bare previously unknown vulnerabilities of senescent cells that may be exploited to promote healthspan.

References

Judith Campisi and Fabrizio d'Adda di Fagagna. Cellular senescence: when bad things happen to good cells. Nature reviews Molecular cell biology, 8(9): 729, 2007.

Abel Soto-Gamez and Marco Demaria. Therapeutic interventions for aging: the case of cellular senescence. Drug discovery today , 22(5):786–795, 2017.

Alejandra Hernandez-Segura, Tristan V de Jong, Simon Melov, Victor Guryev, Judith Campisi, and Marco Demaria. Unmasking transcriptional heterogeneity in senescent cells. Current Biology , 27(17):2652–2660, 2017.

Mitra Mojtahedi, Alexander Skupin, Joseph Zhou, Ivan G Castano, Re-becca YY Leong-Quong, Hannah Chang, Kalliopi Trachana, Alessandro Giuliani, and Sui Huang. Cell fate decision as high-dimensional critical state transition.PLoS biology, 14(12):e2000640, 2016.

Masa Tsuchiya, Alessandro Giuliani, Midori Hashimoto, Jekaterina Erenpreisa, and Kenichi Yoshikawa. Emergent self-organized criticality in gene expression dynamics: Temporal development of global phase transition revealed in a cancer cell line. PloS one, 10(6):e0128565, 2015

Ruiqi Wang, Chunguang Li, Luonan Chen, and Kazuyuki Aihara. Modeling and analyzing biological oscillations in molecular networks. Proceedings of the IEEE, 96(8):1361–1385, 2008

Marten Scheffer, Stephen R Carpenter, Timothy M Lenton, Jordi Bas-compte, William Brock, Vasilis Dakos, Johan Van de Koppel, Ingrid A Van de Leemput, Simon A Levin, Egbert H Van Nes, et al. Anticipating critical transitions.science, 338(6105):344–348, 2012.

Niall M Mangan, J Nathan Kutz, Steven L Brunton, and Joshua L Proctor. Model selection for dynamical systems via sparse regression and information criteria. Proc. R. Soc. A, 473(2204):20170009, 2017.

Gal Berkooz, Philip Holmes, and John L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. Annual review of fluid mechanics, 25(1):539–575, 1993.

Keywords: systems biology, cellular senescence, mathematical modelling, dynamical system, transcriptomics

Allele-specific analysis of the effect of non-coding SNPs on gene expression during Drosophila embryonic development.

Swann Floc'hlay * [1], Bingqing Zhao , David Garfield , Emily Wong , Morgane Thomas-Chollier [1], Denis Thieffry [1], Eileen Furlong [2]

[1] Computational Systems Biology - IBENS (IBENS) – École normale supérieure [ENS] - Paris, Inserm : U1024, CNRS : UMR8197 – 46, rue d'Ulm, 75005 Paris, France

[2] European Molecular Biology Laboratory (EMBL) – Meyerhofstraße 1, 69117 Heidelberg, Germany, Germany

Introduction

Precise regulation of gene expression is essential for almost all biological processes, and a key driving force of development, evolution and disease. The production of messenger RNA is regulated via communication between enhancers, the *cis*-regulatory elements that recruit transcription factors, and the gene's promoter, which recruits the basal transcriptional machinery.

Recent high-throughput sequencing studies between individuals of a given species have revealed extensive variation in gene expression, as a consequence of segregating genetic variation within the population. Most of this regulatory genetic variation is in non-coding DNA, presumably disrupting the function of enhancer elements. However, understanding and predicting how genetic variants disrupts transcriptional regulation remains very poorly understood.

This project aims to get a mechanistic understanding of how natural genetic variation affects multiple layers of transcriptional regulation, using hybrid embryos of genetically distinct *Drosophila* lines isolated from a wild population (MacKay et al., 2012). The use of hybrid individuals offers a powerful approach to dissect *cis* versus *trans*-regulatory mutations, by obtaining allele specific (AS) information. AS measures have the advantages of being biologically interpretable, independent of expression level, applicable to any types of NGS data and offering an easy way to compare, in the same cellular environment, the impact of two different genotypes on a given regulatory layer.

However, working with AS data also introduces a number of interesting bioinformatics challenges, including the control for mapping bias and genotyping errors.

Materials and methods

Experimental methods

The Furlong lab (EMBL, Heidelberg) has generated an F1 embryo collection from 8 different intra-species crosses, at a scale that is sufficient to assay multiple steps of transcription from

*Speaker

the same pool of embryos. Each F1 lines has been sampled at three crucial windows of embryogenesis, corresponding to the events of cell specification (2-4h), cell differentiation (6-8h) and gastrulation (10-12h). Experiments of mRNA-seq, ATAC-seq and histone ChIP-seq (targeting H3K27ac and H3K4me1) have been performed twice on each sample, resulting in a final dataset of approximately 200 samples.

Bioinformatics methods

Most of the bioinformatics pipeline of this project has been realised using the workflow management system Snakemake (Ḱoster and Rahmann, 2012).

In order to correct for potential mapping bias arising from the presence of allelic differences in the obtained reads, we used the ”parental-genomes” strategy, which consists in simultaneously mapping the reads on both parental genomes before assigning them to their parent of origin.

To further correct for mapping bias, we also generated a set of blacklisted genomic, based on the differences in coverage between the mapping of simulated transcriptomic and genomic reads on each of the two parental genomes.

In order to correct the allele-specific ratios for the presence of maternally deposited transcripts, we generated a list of maternally transcribed genes based on the results of mRNA-seq experiments on unfertilised eggs from the Furlong lab. We significantly detect 6,795 genes that we later exclude from the analysis of the mRNA-seq samples.

Furthermore, to correct for genotyping errors we used the results of the sequencing of the genomic DNA (gDNA) of each F1 lines from the collection. We performed a binomial test for imbalance at each SNP position for the AS measure of gDNA, and excluded from the following analysis any SNP showing significant imbalance in the AS ratio (p-value threshold of 5% with FDR correction, requiring a minimum coverage of 20 reads). We discarded on average 10,000 SNP for each cross.

Results

Bioinformatic results

By applying the correction for maternally deposited transcripts, we successfully managed to centre the distribution of AS measures for RNA-seq samples at 0.5, corresponding to a balance state of AS expression (an AS of 1 or 0 corresponding to a complete imbalance toward the mother or the father respectively). This is especially sticking in the case of early mRNA-seq samples, where the bias of AS toward the mother was the highest.

Moreover, the correction for genotyping error based on gDNA data also seems to improve the quality of our measures. Indeed, removing the SNPs imbalanced at the genomic DNA level has the effect to reduce the variability of the AS measures and also remove the number of features (genes, ATAC or ChIP peaks) showing an extreme level or imbalance close to 0 or 1.

Experimental results

Using the AS measures from the gDNA data, we obtained a mean AS ratio specific to the chromosome X of 0.66, which is consistent to our expectation in case of a pool of embryo with an equal proportion of males and females, as only maternal alleles are present in the only chromosome X of the male embryos.

However, the mean AS ratio obtained in RNA-seq and ATAC-seq samples specific to the chromosome X is varying between 0.66 and 0.75 where one would have expected it to be close to 0.75 due to the dosage compensation taking place and increasing the expression of the chromosome X in male embryos. We can hypothesise that this measure, lower than expected, can come from the presence of an incomplete dosage compensation at these stages of development.

Conclusion

Our different tools for correcting the possible bias inherent to allele-specific mapping have allowed us to improve our measure of AS expression all our samples. We have now reliable AS measures from different genotypes, developmental stages and regulatory steps that we can further integrate together to decipher possible links across various dimensions.

The next aim of this project thus consists in the integration of the various levels of regulation, to allow us to disentangle the influence of genetic variation on transcriptional regulation and potentially highlight novel interactions occurring during embryonic development.

References

K´oster, J., and Rahmann, S. (2012). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522.
MacKay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature* 482, 173–178.

Keywords: Allele, specific, F1 crosses, mappability bias, ATAC, seq, RNA, seq

Synteny-guided gene tree correction accounting for whole-genome duplication

Elise Parey * [1], Camille Berthelot [2,3], Hugues Roest-Crollius [4]

[1] Institut de Biologie de l'Ecole Normale Supérieure (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris, CNRS : UMR8197, Institut National de la Santé et de la Recherche Médicale - INSERM : U1024 – 46 rue d'Ulm, Paris F-75005, France

[2] Institut de Biologie de l'Ecole Normale Superieure (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris, CNRS : UMR8197, Institut National de la Santé et de la Recherche Médicale - INSERM – 46 rue d'Ulm, Paris F-75005, France

[3] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI) – Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK, United Kingdom

[4] Institut de Biologie de l'Ecole Normale Supérieure (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris, CNRS : UMR8197, Institut National de la Santé et de la Recherche Médicale - INSERM : U1024 – 46 rue d'Ulm, Paris F-75005, France

Whole genome duplications (WGD) are rare but dramatic events in vertebrate evolution. They generate redundant gene copies, potentially available for the evolution of new functions. For instance, numerous human gene families originate from two whole genome duplications at the origin of vertebrates. Yet, the evolutionary mechanisms affecting genomes after a whole genome duplication are still poorly understood. We study the teleost-specific WGD event that occurred at the root of the teleost fish species tree, dated 320 Mya. This duplication was followed by rapid species radiation, where the genomes of the descending species have retained variable fractions of duplicated genes. The impressive diversity of the teleost clade, making up for half of the extant vertebrate species, renders it a dataset of particular interest to study the genomic consequences of WGD.

One crucial step towards understanding the evolutionary fates of duplicated genes is to correctly identify orthologs and paralogs between species. Orthologs are descended from the same gene in the last common ancestor, while paralogs descend from different ancestral duplicates. This problem is particularly acute in teleosts because after speciation, redundant genes may have been differentially lost in descendant species, causing errors in homology relationships in phylogenetic gene trees. Here, we propose to use specific patterns of synteny conservation to correct such errors and establish high-resolution orthology and paralogy maps in teleost genomes. Our strategy relies on comparing a non-duplicated outgroup fish genome, the spotted gar, with post-WGD duplicated genomes, thus revealing patterns of 'Double-Conserved Synteny' (DCS). We next use these results to sort orthologous and paralogous duplicated segments between fish genomes using deletion patterns of redundant gene copies. The underlying rationale is that duplicated segments derived from the same ancestral post-duplication chromosome share a common ancestry and are likely more similar in their patterns of gene retention and loss. The similarity between duplicated segments is assessed taking into consideration the pattern of absence/presence of gene copies and the proportion of genes pre-annotated as orthologs in gene trees. Repeating a number of such pairwise comparisons between duplicated genomes allows to define groups of orthologous

*Speaker

duplicated genes across many teleost species.
In a next step, the identified groups of orthologous genes are used to derive orthology and paralogy constraints for duplicated genes in gene trees, so that their evolutionary scenario is consistent with WGD. These constraints are represented as a multifurcated constrained tree topology, resolved by maximum likelihood optimization to best fit the sequence data. This approach uncovers many errors where paralogy is confused for orthology, allowing us to correct gene tree topologies where the data are explained equally well (or better) by the optimized tree. We are developing the method on the ten teleost genomes currently present in the Ensembl Compara database. We will then apply it to a wide range of teleost species to deliver a rigorous framework to study the molecular evolution and the function of duplicated genes following a whole genome duplication.

Keywords: Whole genome duplication, genome evolution, gene tree correction

Phylogenetic study of Macrophage Migration Inhibitory Factor (MIF) cytokines and analysis of their role in host - parasite interactions

Claire Michelet * [1], Christine Coustau† [2], Harald Keller‡ [2]

[1] Institut Sophia Agrobiotech – Centre national de la recherche scientifique - CNRS (France), Institut national de la recherche agronomique (INRA) : UMR1355, Université Côte d'Azur (UCA) – 400, route des chappes 06410 Sophia Antipolis, France

[2] Institut Sophia Agrobiotech – UMR1355 INRA-CNRS-Université Nice-Sophia Antipolis – 400, Routes des Chappes 06410 Sophia Antipolis, France

MIF (Macrophage Migration Inhibitory Factor) crucial cytokines of the immune system of vertebrates. Some parasites, such as Filarial worms and ticks, secrete MIF proteins into their vertebrate host, where they modulate the immune response. Recently, we showed that a MIF protein is produced in the salivary glands of aphids and secreted during feeding. The protein interferes with the plant immune system and is required by aphids to establish an interaction with the host plant (Naessens et al., 2015). Plants, too, possess genes encoding MIF proteins (Panstruga et al., 2015), but their function is yet unknown.

The presence of MIF proteins in evolutionary distant organisms and their versatile roles in both host immunity and parasitism raise many questions about the evolution of this protein family. Here, we present a phylogenetic reconstruction of MIF proteins from animal and vegetal kingdoms. The evolutionary history of MIF is discussed with regards to the biology of species (parasitic or free living species) and suggest the existence of differential selection pressures across animal and vegetal phyla.

Keywords: MIF, phylogeny reconstruction, structure reconstruction

*Speaker

†Corresponding author: christine.coustau@inra.fr

‡Corresponding author: harald.keller@inra.fr

New 10x Genomics and SMART-seq2 workflows for single-cell RNA-seq data analysis using Eoulsan

Geoffray Brelurut *† [1,2], Nathalie Lehmann * ‡ [1], Hatim El Jazouli * § [1,2], Céline Hernandez [1], Morgane Thomas-Cholier [1], Denis Thieffry [1], Stéphane Le Crom [1,3], Laurent Jourdren *

1

[1] Institut de biologie de l'Ecole normale supérieure (IBENS) – École normale supérieure [ENS] - Paris, CNRS : UMR8197, Institut National de la Santé et de la Recherche Médicale – 46 rue d'Ulm, 75005 Paris, France

[2] Master Bioinformatique, Normandie Université (UNIROUEN) – Université de Rouen Normandie – 1, rue Thomas Becket 76821 Mont-Saint-Aignan Cedex, France

[3] Institut de Biologie Paris Seine (EPS - IBPS)) – Sorbonne Université UPMC Paris VI – 7-9 quai Saint Bernard 75005 Paris, France

The recent refinement of single-cell RNA sequencing (scRNA-seq) protocols offers a more sensitive and precise means of probing uncharted transcriptional landscapes, which greatly helps unveiling cell-to-cell heterogeneities (*e.g.*, discovery of novel cell types and marker genes, reconstruction of cell lineages). In most cases, these methods require cell isolation and lysis, reverse transcription, and complementary DNA amplification before sequencing either the full transcript (*e.g.*, SMART-seq2) or the 3' end along with a single-molecule tag (*e.g.*, Drop-seq) [1]. Processing data obtained by these protocols, from read filtering to expression counting, requires specific workflows, while the additional analyses may be conducted with common tools. Therefore, we present two new workflows dedicated to Drop-seq, as used by 10x Genomics company, and SMART-seq2 data processing with a common part for advanced data analyses.

Eoulsan [2], a modular workflow engine, provides a reliable (automated functional tests) and open-source (a git repository is freely available [3]) framework for running these workflows and reproducing them. It is an alternative to black-box software or highly customized pipelines. The initial processing steps can take advantage of parallel computing, as a panel of popular job schedulers are supported by *Eoulsan* (*Hadoop*, *TORQUE*, or *HTCondor*) *Eoulsan* also uses the *Docker-Galaxy* layout to integrate the different modules needed to process and analyse full transcript and UMIs sequence data, in addition to the *Java* backbone. The experimental design of an *Eoulsan* workflow is stored in a text file, while the parameters and the workflow steps are listed in another XML file, ensuring flexibility and traceability. This approach allows to swiftly resume large analyses upon trouble-shooting, and guarantees reproducibility. A full documen-

[*] Speaker

[†] Corresponding author: brelurut@biologie.ens.fr

[‡] Corresponding author: lehmann@biologie.ens.fr

[§] Corresponding author: eljazoul@biologie.ens.fr tation (including a quickstart guide) will be available shortly on GitHub.

The initial processing steps provide a full workflow including reads quality checking (with FastQC [12]), reads filtering, mapping to a reference genome (with STAR [13] or bowtie [14]), alignments quality checking, expression counting based on reads or on UMIs (with HT-seq [15] or featureCounts [16]), and a MultiQC [17] report. For 10x Genomics protocol, additional steps are available: (i) cell identification and filtering (based on the distribution of cell barcodes), and (ii) UMI quality processing to get rid of sequencing errors and PCR amplification biases (with network-based methods), both based on the open source UMI-tools software package [18].

Various downstream analysis tools are available for differential gene expression (*SCDE* [4]*)*, cell clustering (*Seurat* [5]) and lineage reconstruction (*Monocle 2* [6]), dimensionality reduction (*ZINB-WaVE* [7]), and soon, network inference [8] and other tools [9][10], depending on the user/developer's choice. Moreover, the workflows will soon utilize the *SingleCellExperiment* [11] R classes to ensure both easier additions of upcoming R modules and interoperability between packages.

In conclusion, *Eoulsan scRNA-seq* pipeline provides an integrated workflow for scRNA-seq data analysis on standalone workstations or on computer clusters. With its modular structure and distributed data processing, it can handle large amounts of data in a reproducible, yet flexible manner.

References

Christoph Ziegenhain et al. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. Molecular Cell, 65: 631-43, 2017.

Laurent Jourdren et al. (2012). Eoulsan: A Cloud Computing-Based Framework Facilitating High Throughput Sequencing Analyses. Bioinformatics, 28: 1542-3.

https://github.com/GenomicParisCentre/eoulsan

Peter V. Kharchenko, Lev Silberstein, and David T. Scadden (2014). Bayesian approach to single-cell differential expression analysis. Nature Methods, 11: 740-2.

Andrew Butler et al. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology.

Cole Trapnell et al. (2014). Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. Nature Biotechnology, 32: 381-6.

Davide Risso et al. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Communications, 9(1), 284.

Zhigang Xue et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature, 500(7464), 593.

Keegan D. Korthauer et al. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biology, 17: 222.

Philipp Angerer et al. (2015). Destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics, 32: 1241-3.

Aaron Lun and Davide Risso (2017). SingleCellExperiment: S4 Classes for Single Cell Data. R package version 1.0.0.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics, 31(2), 166–169. http://doi.org/10.1093/bioinformatics/btu638

Liao Y, Smyth GK and Shi W (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923-30.

Ewels, P., Magnusson, M., Lundin, S., & Ḱaller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047–3048.

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Research, 27(3), 491–499

Keywords: sequencing, singlecell, scRNAseq, SMARTseq2, DROPSeq, Software Development

Transcriptomic analysis of the epipelagic copepod Oithona nana (Crustacea; Cyclopoida) developmental biology

Kevin Sugier * [1], Laso-Jadart Romuald [1], Majda Arif [1], Emmanuelle Petit [2], Marc Wessener [2], Julie Poulain [1], Karine Labadie [2], Jean-Louis Jamet [3], Patrick Wincker [1], Mohammed-Amin Madoui [1]

[1] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay – CEA, CNRS, Université Paris-Saclay – Evry, France
[2] Institut François Jacob, Genoscope – Commissariat à l'Energie Atomique (CEA) – Evry, France
[3] Université de Toulon, Aix Marseille Universités, CNRS/INSU, IRD, MIO UM 110 – CNRS : UMR110, Mediterranean Institut of Oceanography – La Garde, France

Copepods are the most abundant animals on Earth and play an essential role in the marine trophic web and biogeochemical cycles. The cosmopolitan genus *Oithona* is described as one of the most numerous copepods. Its life cycle is completed within 2-3 weeks and is characterized by five stages: eggs, nauplii, copepodites, adults (female and male). Despite its ecological importance, the molecular mechanisms enabling *Oithona* development and sexual differentiation are unknown. Therefore, we investigated its developmental biology through genomic and transcriptomic analyses.

Total mRNA from *Oithona nana* individuals at the five different developmental stages were extracted and sequenced using the Illumina technology. Based on the *O. nana* genome reference, we identified 1,233 (8%) genes differentially expressed, whose 618 were stage-specific (log2(FoldChange)> 1). Among the 81 potential LNR (Lin12-Notch Repeat) coding genes present in the *O. nana* genome, 31 (38%) were up-regulated in at least one of the five developmental stages, whose 20 (64%) were male-specific. Using *WGCNA*, 33 gene modules were identified (p-value≤10-3) and characterized by their possible enrichment in *GO* terms, *Pfam* domains, *KEGG* process, stage-specific genes, LNR coding genes and genes under natural selection. Among the 16 modules showing enrichment, one module of 603 genes was characterized with enrichment in male-specific genes, trypsin genes, genes under natural selection and LNR coding-genes.

These results highlight the important role of the LNR in the developmental and sexual differentiation of *Oithona* and thus are good candidates for protein-protein interaction screening and functional analysis.

Keywords: Copepoda, Oithona, development, transcriptomic analysis

[*] Speaker

Systems biology analysis of interaction and regulation networks in siRNA high throughput screenings.

Claire Rioualen [1], Quentin Da Costa [1], Bernard Chetrit [1], Emmanuelle Charafe-Jauffret [1], Christophe Ginestier [1], Ghislain Bidaut * [1]

[1] Centre de Recherche en Cancérologie de Marseille (CRCM) – Aix Marseille Université : UM105, Institut Paoli-Calmettes : UMR7258, Institut National de la Santé et de la Recherche Médicale : U1068, Centre National de la Recherche Scientifique : UMR7258 – 27 bd Leï Roure, BP 30005913273 Marseille Cedex 09, France

In vitro functional studies using RNA interference (RNAi) screening libraries have recently dramatically improved in throughput speed, quality and genomic coverage with the advent of powerful biochemical methods for perturbing genes transcriptional mechanisms. High-throughput RNAi screenings (HTS) allow quantifying the impact of the deletion of each gene in any particular function, from virus-host interactions to cell differentiation. However, there has been less development for functional analysis and systems biology tools dedicated to RNAi analyses in regards to other fields, such as microarray or NGS. We developed HTS-Net, a network-based analysis program (Rioualen et al., 2017), with the goal to identify gene regulatory modules impacted in high-throughput screenings, by integrating transcription factors-target genes interaction data (regulome) and protein-protein interaction networks (interactome) on top of screening z-scores.
HTS-Net works by discovering subnetworks using a search heuristics that works in parallel on a Protein-Protein Interaction network (Garcia et al., 2012) and a regulation network (TF-DNA interaction). In practice, we superimpose RNAi-based gene scores on the interaction and the regulation maps separately. Each one of them is then searched for high scoring areas. The identified regions of interest, the so-called subnetworks, are extracted and reported. After detection on individual networks, we merge the obtained modules to form meta-subnetworks that incorporate regulation information and PPIs. This approach allows reprioritizing genes by replacing hits into their biological context, which includes their physical interactors and their regulators. HTS-Net produces complete HTML reports for subnetwork exploration and analysis.

In order to evaluate the HTS-Net algorithm, we applied it on three RNAi screen studies. The first study is about discovering genes that sign for human embryonic stem cell identity (hESC) (Chia et al., 2010), the second study is about identifying Cancer Stem Cell regulators (Wolf et al., 2013), and the third study aimed at discovering host-HCV interactors (Tai et al., 2009). We performed our analysis using state-of-the-art PPI and regulation databases. In each analysis, we reported newly found markers, GO enrichments and comparison with the original analyses.

HTS-Net proves better performance than simple gene rankings by z-scores, by re-prioritizing

*Speaker

genes and replacing them in their biological context, as shown by the three studies that we reanalyzed. Formatted input data for the three studied datasets, source code and web site for testing the system are available from the companion web site at http://htsnet.marseille.inserm.fr/. We also compared HTS-Net with two network analysis algorithms, including CARD, a software dedicated to RNAi screening functional analysis (Dutta et al., 2016), and hotnet2, a general-purpose gene network analysis program (Leiserson et al., 2015)

References

Chia, N.-Y., Chan, Y.-S., Feng, B., Lu, X., Orlov, Y.L., Moreau, D., Kumar, P., Yang, L., Jiang, J., Lau, M.-S., et al. (2010). A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature *468*, 316–320.

Dutta, B., Azhir, A., Merino, L.-H., Guo, Y., Revanur, S., Madhamshettiwar, P.B., Germain, R.N., Smith, J.A., Simpson, K.J., Martin, S.E., et al. (2016). An interactive web-based application for Comprehensive Analysis of RNAi-screen Data. Nat. Commun. *7*, 10578.

Garcia, M., Millat-Carus, R., Bertucci, F., Finetti, P., Birnbaum, D., and Bidaut, G. (2012). Interactome-transcriptome integration for predicting distant metastasis in breast cancer. Bioinforma. Oxf. Engl. *28*, 672–678.

Leiserson, M.D.M., Vandin, F., Wu, H.-T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet. *47*, 106–114.

Rioualen, C., Da Costa, Q., Chetrit, B., Charafe-Jauffret, E., Ginestier, C., and Bidaut, G. (2017). HTS-Net: An integrated regulome-interactome approach for establishing network regulation models in high-throughput screenings. PloS One *12*, e0185400.

Tai, A.W., Benita, Y., Peng, L.F., Kim, S.-S., Sakamoto, N., Xavier, R.J., and Chung, R.T. (2009). A Functional Genomic Screen Identifies Cellular Cofactors of Hepatitis C Virus Replication. Cell Host Microbe *5*, 298–307.
Wolf, J., Dewi, D.L., Fredebohm, J., Ḿuller-Decker, K., Flechtenmacher, C., Hoheisel, J.D., and Boettcher, M. (2013). A mammosphere formation RNAi screen reveals that ATG4A promotes a breast cancer stem-like phenotype. Breast Cancer Res. BCR *15*, R109.

Keywords: Systems biology, regulation, siRNA screenings, gene networks, interactome, regulome

Search of therapeutic targets in metabolic pathways of TGF-β using graph coloring approaches

Maxime Folschette *† [1,2], Anne Siegel [2], Vincent Legagneux [1], Carito Guziolowski [3], Nathalie Théret [1,2]

[1] Institut de recherche, santé, environnement et travail (Irset) – Université d'Angers, Universite de Rennes 1, École des Hautes Études en Santé Publique [EHESP], Institut National de la Santé et de la Recherche Médicale : U1085, Structure Fédérative de Recherche en Biologie et Santé de Rennes – 263 avenue Général Leclerc 35042 Rennes Cedex, France
[2] Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) – Universite de Rennes 1, Institut National des Sciences Appliquées - Rennes, Université de Bretagne Sud, École normale supérieure - Rennes, Institut National de Recherche en Informatique et en Automatique, CentraleSupélec, Centre National de la Recherche Scientifique : UMR6074, IMT Atlantique Bretagne-Pays de la Loire – Avenue du général LeclercCampus de Beaulieu 35042 RENNES CEDEX, France
[3] Laboratoire des Sciences du Numérique de Nantes (LS2N) – Université de Nantes, Ecole Centrale de Nantes, Centre National de la Recherche Scientifique : UMR6004, IMT Atlantique Bretagne-Pays de la Loire – Université de Nantes – faculté des Sciences et Techniques (FST)2 Chemin de la HoussinièreBP 92208, 44322 Nantes Cedex 3, France

Hepatocellular carcinoma is the most widespread type of liver cancer. It is difficult to treat and the survival rate if very low. The Transforming Growth Factor β (TGF-β) has been identified as a key component of the tumor propagation, in particular because it induces the Epithelial-Mesenchymal Transition (EMT) of tumor cells. This EMT consists in the transformation of epithelial cells, which are fixed and not very invasive, into mesenchymal cells, having the ability to reshape and move inside the extra-cellular matrix. This transformation obviously marks the passage to a very invasive form of cancer with an increased creation of metastasis.

In this context, the research project I am involved in focuses on the numerous pathways that are involved in the regulation of the EMT, which itself depends on the TGF-β. It can be summarized in four steps:
1) A first step consists in analyzing gene expression data of cancer samples, provided by ICGC [1], in order to point to genes of interest. We used a signature taken from GSEA [2] and consisting of a list of 200 genes that are known to be over-expressed during and after the occurrence of the EMT. This signature allowed a clustering analysis in order to determine two groups of samples (expressing this signature or not) and run a differential analysis on these two groups, that is, comparing the average expression of each gene. Some genes were clearly over- or under-expressed after the occurrence of the EMT. Over the 16'282 genes featured by ICGC, we selected 821 up-regulated genes and 89 down-regulated genes over these criteria: adjusted P-value $<$ 10ˆ–5 and fold-change $>$ 2 (resp. fold-change $<$ –2).

[*] Speaker
[†] Corresponding author: Maxime.Folschette@irisa.fr

2) Starting from this list of notably over- and under-expressed genes given by the previous step, we extracted their direct and indirect upstream regulators (transcription factor relations) from Pathway Commons [3], which is an aggregation of 25 pathway databases related to the human organism. For this, we used the tool Bravo [4], which permits to query Pathway Commons. The result is an oriented SIF graph containing 1197 nodes (representing genes or other components) and 10551 edges (representing regulations). Because of the high heterogeneity of the contents of Pathway Commons [5], some post-processing is also required, such as the fusion of duplicate nodes.
3) The first step of this project provided two lists of over-expressed (+) and under-expressed (−) genes, which gives a partial coloring of the nodes on the network obtained from the second step. Starting from this knowledge, we aim at propagating this coloring to the rest of the nodes using consistency rules, such as: ”the over- or under-expression of a gene must be explained by at least one predecessor”. Such consistency rules are already defined in the tool Iggy [6] which can compute all admissible colorings (in practice, there are many) and output the nodes having always the same coloring, which are called predictions. Among all nodes in the graph, 61 are predicted as over-expressed or under-expressed, and are a first interesting hint towards therapeutic targets.
4) Nevertheless, finding more accurate targets would benefit from understanding not only the predictions, but their causes. Several ways are possible to accomplish this. One of them would consist in computing the key regulators, that are, sets of nodes which are minimal, so that their coloring allows to minimize the gap with the gene expression data, and the complement of their coloring allows to maximize this difference. Other ways include dynamical analysis on the graph.

Hudson, T. J. et al. (2010). International network of cancer genome projects. Nature, 464. http://icgc.org/

Subramanian, A. et al. (2005). Gene Set Enrichment Analysis : A knowledge-based approach for interpreting genome-wide expression profiles. Proc. of the Nat. Ac. of Sci., 102(43). http://software.broadinstitute.org/gsea/

Cerami, E. G. et al. (2010). Pathway Commons, a web resource for biological pathway data. Nucleic acids research, 39. http://www.Pathway Commons.org/

Lefebvre, M. et al. (2017). Regulatory and signaling network assembly through linked open data. In Journées Ouvertes en Biologie, Informatique et Mathématiques. Demo paper. https://github.com/symetric-group/bionets-demo

Coquet, J. (2017). Étude exhaustive de voies de signalisation de grande taille par clustering des trajectoires et caractérisation par analyse sémantique. PhD thesis, Université de Rennes 1.

Thiele, S. et al. (2015). Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies. BMC Bioinformatics, 16(1). http://bioasp.github.io/iggy/

Keywords: TGFβ, cancer, hepatocellular carcinoma, graph coloring, differential analysis, graph reconstruction

The reconstruction of the metabolic network of the Tisochrysis lutea microalgae thanks to the AuReMe workspace leads to analysis of carnosine and beta-carotene pathways

Jeanne Got * [1], Gregory Carrier† [2], Anne Siegel‡ [3]

[1] DYLISS (INRIA - IRISA) – CNRS : UMR6074, L'Institut National de Recherche en Informatique et e n Automatique (INRIA) – Campus de Beaulieu 35042 Rennes cedex, France
[2] Physiologie et biotechnologie des Algues (PBA) – Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) – Rue de l'Ile d'Yeu BP 21105 Nantes cedex 3, France
[3] DYLISS (INRIA - IRISA) – INRIA, Universite de Rennes 1, CNRS : UMR6074 – Campus de Beaulieu 35042 Rennes cedex, France

INTRODUCTION

A main challenge of the era of massive genome sequencing is to transform sequences into biological knowledge. Thanks to the reconstruction of metabolic networks, we start from genome sequences, and we describe the majority of the biochemical reactions of a studied species. In 2010, Thiele and Palsson described a general protocol enabling the reconstruction of high-quality metabolic networks (Thiele and Palsson, 2010).

For that matter, microalgae have a big potential for economic benefit because applications are envisaged in very large domains, from human health to biofuel production. *Tisochrysis lutea* (Bendif et al. 2013) is a eukaryotic microalga that belongs to the Haptophyta phylum, and more precisely to the Isochrysidaceae family. From a biological point of view, Haptophytes have several metabolic characteristics, two of which we want to pay your attention especially to two of them. First, the two metabolites carnosine and glutathione have found in several microalgae (Holdt and Kraan, 2011); second, many microalgae produce high levels in beta-carotene (Takaichi, 2011 and Guedes et al., 2011).

In this work, we reconstructed the genome-scale metabolic model (GSM) of *Tisochrysis lutea* thanks to the workspace AuReMe (AUtomatic Reconstruction of Metabolic networks, Aite et al., 2018) that has been developed in our team. Then we present two relevant pathways of *T. lutea*: carnosine biosynthesis and beta-carotene biosynthesis.

METHOD

AuReMe is a workspace dedicated to the generation of GSMs. This workflow based on the Docker technology (https://docker.com), gathers academic-free tools and databases. It designs

*Speaker

†Corresponding author: gregory.carrier@ifremer.fr

‡Corresponding author: anne.siegel@irisa.fr

reconstruction pipelines that are flexible and can suit various available data sources while storing metadata. It can follow four major steps of reconstruction processes: annotation or orthology-based modelings, gap-filling and manual curation. In addition, AuReMe supports most processes of the Thiele and Palsson protocol (Thiele and Palsson, 2010) by proposing tools and methods that facilitate analysis and storing of the results at each step related to experiments or exploration of literature. The following software was installed in the AuReMe environment: Meneco (Prigent et al., 2017) for gap-filling, the MeneTools package (MEtabolic NEtwork TOpological toOLS), and CobraPy (Ebrahim et al., 2013), a package for calculating the Flux Balance Analysis (FBA). FBA is a mathematical approach for analyzing the flow of metabolites through a metabolic network. A PADMet Python package (MEtabolic NEtwork TOpological toOLS), and CobraPy (Ebrahim et al., 2013), a package for calculating the Flux Balance Analysis (FBA). FBA is a mathematical approach for analyzing the flow of metabolites through a metabolic network. A PADMet Python package (Python library for hAndling metaData of METabolism), as a data manager, stores all the necessary information about used methods and how tools are chained in order to facilitate the network reconstruction reproducibility (Aite et al., 2018). Futhermore, AuReMe encompasses these four tools for finding orthologuous: Blastall (Altschul et al., 1990), Inparanoid (Remm et al., 2001), OrthoMCL (Li et al., 2003), and Pantograph (Loira et al., 2015). The MediaWiki technologies (https://mediawiki.org) were also encapsulated in the Docker container. MediaWiki is useful in order to produce the representation of the metabolic model through local wiki webpages. The AuReMe environment supplies various databases like MetaCyc (Caspi et al., 2016), BiGG (King et al., 2016), ModelSEED (Henry et al., 2010), and the MetaNetX dictionary (Moretti et al., 2016).

To reconstruct and analyze the metabolic network of *T. lutea*, we employed all of the aforementioned software from AuReME. We also utilized Pathway-Tools (Karp et al., 2002) that created a first draft GSM from an annotated genome and the Blast Reciprocal Best Hit (RBH) method to retrieve some reactions.

RESULTS

We reconstructed the metabolic network of *T. lutea* by using the workspace AuReMe. We employed the Pathway-Tools (Karp et al., 2002) to create a first draft GSM. Then, four additional metabolic draft networks were generated by searching orthologuous genes in other GSMs with Pantograph (Loira et al., 2015). The GSMs of *Arabidospsis thaliana*, a plant and one a the most studied organism (de Oliveira Dal'Molin et al., 2010), *Chlamydomonas reinhardtii*, a well-known green microalga (Imam et al., 2015), *Ectocarpus siliculosus*, a brown alga model organism (Prigent et al., 2014), and *Synechocystis* sp. Pcc 6803, a very well studied cyanobacterium (Knoop et al., 2013) were used as template models.

We also had a small-scale network of primary metabolism of T. lutea (281 metabolites and 261 reactions), enabling growth simulations through FBA at our disposal.

All the previously described draft networks (i.e: the annotated one, the four ones from the orthology, and the small-scale network) were reconstructed with distinct databases, with different metabolite and reaction identifiers. In order to merge them, we employed the PADMet package (Aite et al., 2018), which mapped these networks to the Metacyc database (Caspi et al., 2016), by combining a systematic use of the MetaNetX dictionary (Moretti et al., 2016) and manual curation. Furthermore, we gap-filling was performed on the resulting network with Meneco (Prigent et al., 2017) and the MeneTools.

The reconstruction of the *T. lutea* genome-scale metabolic encompasses 2.728 genes related to 2.799 reactions and 2.747 metabolites. According to our Flux Balance Analysis results, the reconstructed network was able to produce biomass. This GSM is explorable on the http://gem-aureme.irisa.fr/tisogem/ website.

We explored the GSM of *T. lutea* by browsing through its wiki to analyze some relevant pathways. Among these pathways, we studied the carnosine and the beta-carotene production.

The analysis of the carnosine pathway strongly suggests that *T. lutea* has the same capability as *Chlamydomonas reinhardtii* to produce carnosine, a specific antioxidant dipeptide consisting of beta-alanine and L-histidine. Beta-alanine is produced through two distinct pathways, including one initiated by aspartate. On the contrary, only one of them was identified in the brown macroalga *Ectocarpus siliculosus.* Interestingly, the missing pathway producing beta-alanine from aspartate was identified in a symbiont of its algal wall: *Candidatus* Phaeomarinobacter *ectocarpi* (Dittami et al., 2014), paving the way to the study of organisms communities at the metabolic level.

Beta-carotene is the precursor of all carotenoids and xanthophylls in plants and algae. The second metabolic pathway we manually examined was the productions of beta-carotene, a pigment known to be produced at high levels in several Haptophytes (Takaichi, 2011). Our analyses confirm that many the upstream pathways from the beta-carotene production are complete in *T. lutea.* We only added three reactions (RXN-11354, RXN-12243, and RXN-12244 based on manual curation) to fill in the MetaCyc pathway PWY-6475 in charge of the trans-lycopene biosynthesis in the GSM of *T. lutea. T. lutea* can also synthesize the beta-carotene via the phytoene and the neurosporene. *C. reinhardtii* produces the beta-carotene too, but the pathway PWY-6475 is not complete in its metabolic network. In fact, four reactions (RXN-11354, RXN-12243, RXN-12244 and, RXN-8042) are missing from the GSM of *C. reinhardtii.* The PWY-6475 seems also to be complete in *E. siliculosus.*

To sum up, the comparison of the several interested organisms at the pathway level, allows us to better understand the biology of our models, and improve our GSMs.

CONCLUSION
AuReMe is a workspace to reconstruct GSMs with ”a la carte” pipelines. Thanks to AuReMe we generated the metabolic network of *Tisochrysis lutea*, a eukaryotic microalga. This GSM encompasses 2.728 genes related to 2.799 reactions and 2.747 metabolites. It is exploitable on the http://gem-aureme.irisa.fr/tisogem/ website. Furthermore, the analyses of two interesting pathways, that carosine biosynthesis and that beta-carotene biosynthesis, allowed us to evaluate the quality of our GSM, because *T. lutea* is able to produce carnosine and beta-carotene.

Keywords: Metabolic network reconstruction, microalga, biological pathway analysis, workflow, non, model species

PSSMSearch: discovery of protein motifs on a proteome-wide scale

Izabella Krystkowiak [1,2], Jean Manguy * [1,2,3], Norman Davey† [1,2]

[1] UCD Conway Institute (UCD) – UCD Conway Institute of Biomolecular and Biomedical Research
University College Dublin, Belfield, Dublin 4, Ireland
[2] UCD School of Medicine Medical Science (UCD) – UCD School of Medicine Medical Science
University College Dublin, Belfield, Dublin 4, Ireland
[3] Food for Health Ireland (FHI) – Food for Health Ireland Science Centre South, UCD, Dublin 4, Ireland

Many functions of a protein are mediated by protein binding motifs (short linear motifs, SLiMs, ELMs or miniMotifs) and modification motifs (moiety attachment/removal, isomerization, cleavage). PSSMSearch uses a set of aligned peptides of a motif of interest to create a position-specific scoring matrix (PSSM) describing the binding determinants of a motif. These peptides can be defined by the user or downloaded from the Eukaryotic Linear Motif (ELM) database. The PSSM can be generated using several scoring methods including PSI-BLAST, MOTIPS, Log Relative Binomial and log odds, and different background amino acid frequencies. Additionally, the PSSM can be modified by a user to add prior knowledge of binding determinants through an interactive PSSM heatmap. PSSMSearch then scans the PSSM against the proteome of a selected species to discover putative novel motif instances. Currently, 70 proteomes are available from all kingdoms including all species of experimental and therapeutic relevance. PSSMSearch returns a table of PSSM matches along with information regarding the statistical significance of the match, the position of the motif in the protein, shared annotations with a selected motif-recognising partner(s) (if specified), motif attributes (such as accessibility, conservation) and overlapping protein feature annotations. In addition to this table, the PSSMSearch framework includes tools to further investigate the properties of the motif including taxonomic range calculations and functional enrichment analyses. PSSM matches can be filtered using multiple rules defined by the user based on accessibility, taxonomic range, localisation, interacting partner or functional annotations. The list of matches and their annotation can then be downloaded for further analysis. PSSMSearch is available at http://slim.ucd.ie/pssmsearch/.

Keywords: protein motif, proteome, web server, position specific scoring matrix

*Speaker

†Corresponding author: norman.davey@ucd.ie

Gravity: a tool to analyze genetic variants in individuals using gene networks

Freddy Cliquet* [1], Alexandre Mathieu [2], Thomas Kergrohen [2], Thomas Rolland [2], Guillaume Dumas [2], Julien Fumey † [2], Richard Delorme [2,3], Benno Schwikowski [1], Thomas Bourgeron [2,4]

[1] Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI) – Institut Pasteur [Paris] – 25-28 rue du docteur Roux, 75724 Paris cedex 15, France
[2] Génétique humaine et Fonctions cognitives - Human Genetics and Cognitive Functions – Institut Pasteur [Paris], Centre National de la Recherche Scientifique : UMR3571 – C3BI / Département de Neuroscience - 25-28, rue du docteur Roux, 75724 Paris cedex 15, France
[3] Hôpital Robert Debré Paris – Hôpital Robert Debré – France
[4] Université Diderot – Université Paris Diderot - Paris 7 – France

In many complex diseases such as psychiatric disorders, a combination of common and rare genetic variants, affecting different cellular pathways, underlies the disease susceptibility. Protein-protein interaction networks are used as a backbone of cellular processes, on which the effect of disrupting mutations can be analyzed. However, a tool to visualize and identify causal mutations within interaction networks and across large cohorts is lacking.
Gravity (for Gene inteRaction Analysis of Variants in Individuals: a Tool for You) provides a visualization of genetic data from SNP array, whole genome and whole exome data in the context of protein-protein interactions. In addition, it allows the user to dynamically filter on sequencing quality or allelic frequency, or to focus on specific pathways of interest. The tool integrates analyses of mutation pattern across full cohorts of patients and their relatives, identifying multiple hits in individuals and providing a precise characterization of the variants.

Gravity takes as an input a Gemini database generated from a variant calling file containing all the variants identified in a cohort. As an output,the visualization of the variants found in an individual or in a family includes the network representation, the full annotation of each variant, *i.e.*frequency, deleteriousness and inheritance patterns, as well as cohort details on other carrying individuals. The network representation, as well as the gene annotation and sequencing parameters, can be exported through Cytoscape functions or tab-separated files.

Gravity helps in the interpretation of the combined effect of multiple mutations within different pathways of interest in the patients, setting the stage for better patient stratification and potentially more precise and personalized therapeutic strategies.

This application has been routinely used on six different cohorts in our lab over the past year (up to 300 individuals with whole-exome sequencing and 170 with whole genome sequencing), and has been tested by more than 30 persons from outside our group, including bioinformaticians, geneticists and clinicians.

*Corresponding author: freddy.cliquet@pasteur.fr

†Speaker

The tool and its documentation can be found at http://gravity.pasteur.fr
The source code is available at https://bitbucket.org/fcliquet/gravity/overview

Keywords: gravity, cytoscape, network, genomics, WGS, visualisation, dataviz

Optimization of RNA-seq differential expression analysis and transcriptome profiling of metabolism and triacylglycerol biosynthesis in a novel halotolerant chlorella species

Michaël Pierrelée * [1], Jin Ho Yun [2,3], Hee-Sik Kim [2]

[1] Institut de Biologie du Développement de Marseille (IBDM) – Aix Marseille Université : UMR7288, Institut National de la Santé et de la Recherche Médicale : UMR7288, Centre National de la Recherche Scientifique : UMR7288 – Case 907 - Parc Scientifique de Luminy 13288 Marseille Cedex 9, France
[2] Korea Research Institute of Bioscience Biotechnology (KRIBB) – 125 Gwahak-ro, Yuseong-gu, Daejeon, South Korea
[3] Korea Advanced Institute of Science and Technology (KAIST) – 291 Daehak-ro, Guseong-dong, Yuseong-gu, Daejeon, South Korea

Algal lipids have gained much attention for its commercial applications in food, cosmetic and bioenergy sectors. However, the production of algal lipid is not yet competitive enough. A *Chlorella* species, HS2, was recently isolated and has strong industrial potential due its oleaginity and halotolerance. Adaptation of HS2 to salt stress at the transcriptome level was explored by RNA-seq. This technology necessitates the differential expression analysis and its current normalization algorithm calls for improvement to overcome intrinsic potential issues. In this study, a new normalization algorithm, SVCD, was adapted to RNA-seq. The results indicated that SVCD identified more differentially expressed genes, but only those directly related to salt stress were selected. Although there was a strong asymmetry in favor of up-regulated transcripts, due to activation of translation, shifts in protein synthesis and degradation were lower and higher, respectively. Up-regulation of synthesis of several sugars and amino acids related to osmoprotection was also observed. In addition, enzymes generating pyruvate, malate and acetyl-CoA are up-regulated, giving precursors to *de novo* triacylglycerol synthesis whose proteins were more expressed.

Keywords: RNA, seq, normalization, differential expression analysis, algae, lipid production, metabolism, biofuel

*Speaker

From gene expression to genomic networks : development of a user-friendly pipeline for the analysis and visualization of RNA-seq data

Ilana Lambert * [1,2], Christine Paysant - Le Roux [2,3], Marjorie Pervent [1], Antoine Le Quere [1], Marc Tauzin [1], Marc Lepetit [1], Marie-Laure Martin-Magniette [2,3,4], Stefano Colella [1]

[1] Laboratoire des symbioses tropicales et méditerranéennes (LSTM) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement : UMR82, Institut National de la Recherche Agronomique, Université de Montpellier, Institut national d'études supérieures agronomiques de Montpellier, Institut de Recherche pour le Développement - IRD (FRANCE) – Campus international de Baillarguet - TA 10 / J - 34398 Montpellier Cedex 5, France

[2] Institute of Plant Sciences Paris Saclay (IPS2) – Institut national de la recherche agronomique (INRA) : UMR1403, CNRS : UMR9213, Université Paris Sud - Paris XI, Université d'Evry-Val d'Essonne, Université Paris Saclay – bâtiment 630, rue Noetzlin 91405 Orsay, France

[3] Institute of Plant Sciences Paris Saclay (IPS2) – Université Paris Diderot - Paris 7, PRES Sorbonne Paris Cité – bâtiment 630, rue Noetzlin 91405 Orsay, France

[4] Unité de recherche Mathématiques et Informatique Appliquées UMR518 (MIA) – AgroParisTech, Institut national de la recherche agronomique (INRA), Université Paris-Saclay – 16 rue Claude Bernard F-75231 Paris Cedex 05, France

RNA-seq data produced using Next-Generation high-throughput Sequencing technology (NGS) is nowadays the preferred method to study gene expression. Differential gene expression analysis across different experimental conditions has been extensively used over the years to gain insight in gene functions through the characterization of their implication in biological processes. The characteristics of RNA-seq data, such as heterogeneity of counts or over dispersion among biological replicates, represent a methodological challenge. Therefore, the large-scale data analyses of RNA-seq data are not straightforward for a biologist because a number of methods to filter and normalize data, as well as different statistical approaches for gene expression analysis, have emerged in the literature in recent years. Furthermore, the implementation of the analysis methods requiring the acquisition of statistical and programming skills in R may represent a difficulty for most biologists.

The objective of the work presented here is the development and implementation of a user-friendly pipeline for the analysis and visualization of RNA-seq data using methods chosen following recommendations in methods comparative evaluation published studies. The user is guided through all the analysis steps, making the pipeline usable by biologists without advanced programming skills and statistical knowledge. The analysis pipeline is composed of pre-existing tools for RNA-seq analyzes, as well as an *ad hoc* R package, combined to perform five main analysis steps:

[*] Speaker

Filtration to discard no-expressed and low-count genes followed by the normalization to correct library-specific biases.

Identification of differential expressed genes among different conditions.

Co-expression analysis to identify genes sharing profiles of expression.

Functional classes enrichment analysis to identify an over-representation of specific annotation terms in the lists of genes of interests.

Integration of the results of the pipeline to generate gene expression networks.

The selection of the methods based on bibliographic studies is the foundation of the development of the analysis pipeline. We chose the data filtering method, based on the work of Rau *et al.* [1] which compared two different approaches: the mean-based filter RPKM (Reads Per KiloBase per Million mapped reads) and the maximum-based filter CPM (Counts Per Million) methods. Maximum-based filters have better sensitivity for genes with low levels of expression and we chose to implement the CPM method in our analysis pipeline.

For the normalization, we based our choice on the work of Dillies *et al.* [2] that compared multiple methods. In this work, the authors showed that the "Relative Log Expression" (RLE) method developed in the package DESeq2 [3] and the "Trimmed Mean of M-value" (TMM) method developed in the package edgeR [4] demonstrate satisfactory behavior in the presence of highly expressed genes. Moreover, they showed that only these methods could maintain a good false-positive rate without loss of power. Between them, one advantage of the TMM method is to be less affected by the proportion of differential expressed genes and that is why TMM was chosen for implementation in our pipeline.

To choose the statistical model, we based our choices mainly on the work of Rigaill *et al.* [5]. In this study, authors made neutral comparisons between negative binomial-based method, generalized linear models and finally linear models on transformed data. Performance analyses based on the p-value distributions, ROC curves and proportion of true and false positive rates show a clear difference of behavior between negative binomial-based methods and the others. Linear models on transformed data or generalized linear models are consequently the most adapted for the differential analysis even in case of multi-factor comparisons as showed also by Lin *et al.* [6]. Since we have chosen the TMM normalization available in the edgeR package, we developed the generalized linear models (GLM) using this package in the pipeline.

Rau *et al.* [7] worked on co-expression analysis concluding that normalized expression profiles are recommended for co-expression analyses of RNA-seq data. They showed that after a transformation of the normalized expression profiles, Gaussian mixture models are suitable for RNA-seq analyzes and provide good identification of the groups of co-expressed genes. These mixture models are preferable to those based on other distributions because they account for per-cluster correlation structures among samples. We will integrate these analyzes into the pipeline using the coseq R package [8] to generate co-expressed gene clusters that will be presented as gene expression networks.

To help the biologist in their further analysis of gene lists, generated with the GLM differential analysis, the co-expression analysis and/or the combination of the two, we included at different stages in the pipeline some visualization graphs (histograms, Venn diagrams, hierarchical clustering, PCA, ...). Furthermore, the pipeline will allow the biologists to perform functional annotations enrichment analysis of the gene lists using hypergeometric tests. While several tools for Gene Ontology enrichment analysis exists as stand-alone or online, a generic tool integrated in an analysis pipeline will be of great interest to use other kind of annotations, including expertise based in-house functional annotations.

We tested the pipeline on several datasets available in the laboratory on *Medicago truncatula*, the genomic model for legume plants. Legume have the capacity to associate with soil nitrogen fixing bacteria (*Rhizobia*) to form specific root endosymbiotic organs called nodules. Symbiotic bacteria fix atmospheric N2, a non-limiting source of nitrogen (N) and provide it to legumes allowing them to overcome soil mineral nitrogen shortage, while obtaining carbon from plant photosynthesis. However nodules are highly sensitive to environmental stresses. Local stresses inhibiting nitrogen fixation are integrated at the whole plant level to generate long distance N-systemic responses that promote root and nodule development in non-stressed soil areas. To discriminate unambiguously in *Medicago truncatula / Rhizobium* symbiotic plants between direct responses associated to local stresses and responses related to N-systemic signalling an experimental split-root system was used. The test dataset that I will present includes 24 samples and 6 different biological conditions. The aim of the study was to characterize the systemic response of the whole plant to N-deficit over the nodule development process (7 days).

We developed a rigorous structure in R programming to allow the users to follow the analyses at each step with tutorials and suggested questions. The results are presented in the form of summary or comparative tables and graphs to facilitate the analysis. The pipeline combines different scripts that are automated allowing easily to use it on multiple and different datasets. In this automation process an R package is being developed to allow writing generalized linear models (GLM) for every possible comparisons.

Gene networks representations are nowadays widely used to bring new knowledge and working hypotheses on gene regulation processes. The combination of differentially expressed genes with co-expression clusters implemented in our pipeline will leads to the generation of potential gene networks. The data generated will be included in novel *ad hoc* database to be implemented for *M. truncatula*, based on the GEM2Net database for co-expression data on biotic and abiotic stress categories in *Arabidopsis thaliana* developed by Zaag *et al.* [9]. Data included in GEM2Net were produced with a similar analysis strategy but not using an automated analysis pipeline. Our pipeline will thus allow performing standardized and automated analyses of existing and novel RNAseq datasets to enrich the novel database for exploration of gene regulations in the legume model *M. truncatula*.

References :

[**1**] Rau A, Gallopin M, Celeux G, Jaffrezic F. (2013) ”Data-based filtering for replicated high-throughput transcriptome sequencing experiments.” *Bioinformatics*, 29 pp. 2146–2152.

[**2**] Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Lalo´e D, Le Gall C, Scha´effer B, Le Crom S, Guedj M, Jaffrézic F, French StatOmique Consortium. (2013) ”A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis.” *Briefings in Bioinformatics*, 14(6) pp. 671–683.

[**3**] Love MI, Huber W and Anders S (2014). ”Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, pp. 550.

[**4**] Robinson MD, McCarthy DJ and Smyth GK (2010). ”edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” Bioinformatics, 26(1) pp.

139-140.

[**5**] Rigaill G, Balzergue S, Brunaud V, Blondet E, Rau A, Rogier O, Caius J, Maugis-Rabusseau C, Soubigou-Taconnat L, Aubourg S, Lurin C, Martin-Magniette ML, Delannoy E. (2018) ”Synthetic data sets for the identification of key ingredients for RNA-seq differential analysis” *Briefings in Bioinformatics*, 19(1) pp.65-67

[**6**] Lin Y, Golovnina K, Chen ZX, Lee HN, Negron YL, Sultana H, Oliver B, Harbison ST. ”Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila Melanogaster.” *BMC Genomics.* 2016;17:28.

[**7**] Rau A and Maugis-Rabusseau C (2017). ”Transformation and model choice for co-expression analysis of RNA-seq data.” Briefings in Bioinformatics.

[**8**] Rau A, Maugis-Rabusseau C, Martin-Magniette M-L, Celeux G. (2015) Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics*, 31(9) pp.1420-7

[**9**] Zaag R., Tamby J. P., Guichard C., Tariq Z., Rigaill G., Delannoy E., Renou JP., Balzergue S., Mary-Huard T., Aubourg S., Martin-Magniette ML., Brunaud V. (2015). ”GEM2Net: from gene expression modeling to -omics networks, a new CATdb module to investigate Arabidopsis thaliana genes involved in stress response.” *Nucleic Acids Res.* 43 D1010–D1017.

Keywords: RNA, seq, genomic networks, pipeline

TOGGLe (Toolbox for Generic NGS Analyses) A framework to quickly build pipelines and to perform large-scale HTS analysis.

Julie Orjuela * [1,2,3,4], Sébastien Ravel *

[1,5], Alexis Dereeper *

[1,2], Ndomassi Tando *

[1,3], François Sabot *

[1,3], Christine Tranchant-Dubreuil *

1,3

[1] South Green Bioinformatics Platform (SG) – Bioversity, CIRAD : UMRAGAP / BGPI / LSTM, Institut de recherche pour le développement [IRD] : UMRDIADE/ IPME – Montpellier, France
[2] UMR - Interactions Plantes Microorganismes Environnement (UMR IPME) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Université de Montpellier, Institut de Recherche pour le Développement : UMR186 – IRD France-Sud 911, avenue Agropolis BP 64501 34394 Montpellier cedex 5, France
[3] UMR DIADE IRD/UM (DIADE) – Université de Montpellier, Institut de Recherche pour le Développement – Centre IRD de Montpellier 911 av Agropolis BP 604501 34394 Montpellier cedex 5, France
[4] Biologie des Organismes et Ecosystèmes Aquatiques (BOREA) – Institut de Recherche pour le Développement – 7, rue Cuvier, CP 32, 75231 Paris Cedex 05, France
[5] Biologie et Génétique des Interactions Plante-Parasite (BGPI) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement : UMR54, Institut National de la Recherche Agronomique : UMR0385, Centre international d´tudes supérieures en sciences agronomiques, Institut national d'études supérieures agronomiques de Montpellier, Centre international d´tudes supérieures en sciences agronomiques, Centre international d´tudes supérieures en sciences agronomiques, Centre international d'études supérieures en sciences agronomiques – CIRADUMR-BGPI TA A-54/K - Campus International de Baillarguet - 34398 Montpellier Cedex 5, France

High throughput sequencing (HTS) data analyses are done every day for biologist and bioinformatics in order to give biological sense of their data. These analyses must to be reproducible, robust and efficient. A generic tool TOGGLe (Toolbox for Generic NGS Analyses) was devel-

*Speaker

oped to allow running simple and complex pipelines without require any programming skills. This workflow manager is friendly to users and transparent for developers. User only need basic Linux commands and to specify freely their favorite software parameters through a simple text file given.
TOGGLe manages, controls, verifies and concatenates every step in your favorite workflow. This workflow manager checks structure and compatibility of steps given in the software parameters file, it builds a workflow and launches it. TOGGLe reports parameters, commands executed, software versions as well as errors if they occur. These informations are kept in logs and reports files. Results are organized in a structured tree of directories. TOGGLe allows compressing or removing intermediate data and it uses scheduler machinery.

TOGGLe integrates a large panel of tools for HTS analyses (demultiplexing, cleaning, trimming, calling, assembly, structural variation detection, transcriptomic ...) and post-analysis (haplotype detection, population structure ...). This workflow manager is highly flexible on the data type, working on sequencing raw data (Illumina, 454 or Pacific Biosciences), as well as on various other formats (e.g. SAM, BAM, VCF). This HTS workflow manager allows running parallel analysis or global, where several samples will be analysed together.

TOGGLe was used on numerous sequencing projects with high number of samples, or/and high depth sequencing. It was shown to be highly adaptable to various biological questions as well as to a large array of computing architecture and data. In this poster, we are going to show how users can enjoy of TOGGLe in their data analysis.

Project home page: http://toggle.southgreen.fr

Code repository: https://github.com/SouthGreenPlatform/TOGGLE

Programming Language: Perl 5.16 of higher (5.24 Recommended)

License: GNU GPLv3/CeCill-C
Paper: Tranchant-Dubreuil, C., Ravel, S., Monat, C., Sarah, G., Diallo, A., Helou, L., ... Sabot, F. (2018). TOGGLe, a flexible framework for easily building complex workflows and performing robust large-scale NGS analyses. BioRxiv. https://doi.org/10.1101/245480

Keywords: NGS, HTS, pipeline, workflow, reproductive, friendly use.

Phage Remote Ortholog Groups (PhROG) : clustering phage proteins using remote homology

Paul Terzian * [1], Robin Mom [1], Clovis Galiez [2], Julien Lossouarn [3], Ariane Toussaint [4], Marie-Agnès Petit [3], François Enault† [1]

[1] Université Clermont Auvergne, CNRS, LMGE, F-63000 Clermont-Ferrand, France – CNRS : UMR6023 – France

[2] Quantitative and Computational Biology Group, Max-Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany – Germany

[3] Institut Micalis, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France. – Institut national de la recherche agronomique (INRA) – France

[4] University Libre de Bruxelles, Génétique et Physiologie Bactérienne (LGPB), 12 rue des Professeurs Jeener et Brachet, 6041 Charleroi, Belgium – Belgium

Viruses are the most abundant biological entities on earth and represent the widest unknown source of genetic diversity. The recent increase in viral sequenced genomes represent a great opportunity to gain new insights into this diversity and consequently urges the development of automatized annotation tools to help functional and comparative analysis. Here, we introduce PhROG (Phage Remote Ortholog Groups), a new method to clusterize viral sequences using remote homology through HMM profile comparisons. Compared to already known protein clustering approaches (POG, EggNOG...), more proteins were clustered using our methodology when applied to a set of 50,000 viral proteins. Furthermore, the average coverage inside each cluster was also greater. This method was then applied to the proteins of the 1,694 reference viruses infecting prokaryotes and 12,498 viral genomes identified in bacterial and archaeal genomic data sets. The resulting clusters were annotated using different strategies and databases, coupled with careful and manual inspection. Finally, we used these clusters to annotate 60,000 viral proteins from different environments and managed to annotate 10% more proteins on average than with a traditional BLASTp comparison against RefSeq. PhROG will be a useful database to better annotate future phage genomes and viral metagenomes.

Keywords: protein clustering, remote homology, phages

*Speaker

†Corresponding author: Francois.ENAULT@uca.fr

CoRegCAD: leveraging networks for metabolic engineering

Pauline Trébulle * [1,2], Jean-Marc Nicaud [3], Mohamed Elati† [1]

[1] Université Lille, CNRS, Centrale Lille, CRIStAL (UMR 9189) – Centre National de la Recherche Scientifique : UMR9189, Université Lille I - Sciences et technologies, Ecole Centrale de Lille – France
[2] MICrobiologie de l'ALImentation au Service de la Santé humaine (MICALIS) – AgroParisTech, Institut national de la recherche agronomique (INRA) : UMR1319 – F-78350 JOUY-EN-JOSAS, France
[3] MICrobiologie de l'ALImentation au Service de la Santé humaine (MICALIS) – Institut National de la Recherche Agronomique : UMR1319, AgroParisTech – F-78350 JOUY-EN-JOSAS, France

Introduction
Bio-design automation (BDA) and biological computer-aided design (BioCAD) tools are crucial for the development of synthetic biology and industrial biotechnology which aim at designing and engineering large, self-adaptive, coupled regulatory and metabolic systems at whole-genome scale for useful purposes in a cost-effective manner.

Although the landscape of BDA and CAD tools has significantly grown for the last few years [1], in particular regarding the design of complex genetic circuit based on characterized part and specification, tools for context-specific and adaptive rational pathway design are yet to be generalized. BioCAD strategies currently available are dependent on a combination of standardized partial characterization and mathematical models for predicting the behavior of the designed device. However, these approaches are limited by the uncertainty inherent to biological systems. This work aims at providing a framework for the design and optimization of pathways and phenotypes of interest in industrial strains. To meet that goal, our team has developped several building blocks integrated together in CoRegCAD in an iterative process from network inference and interrogation [2] of the strain regulatory process to the integration of genome architecture when re-factoring chromosomes [3]

In this study, we propose to combine regulatory and metabolic network to:

- Identify the best constructions to improve the production yield in context-specific conditions

- Highlight new regulatory elements of interest for further characterization and integration in parts libraries.

This work will be demonstrated on *Yarrowia lipolytica*, a chassis of industrial interest for which standardized Golden Gate modular cloning strategy has been developed [4].

Materials and Methods

[*] Speaker
[†] Corresponding author: mohamed.elati@univ-lille.fr

CoRegCAD framework includes several tools working together. From a large dataset, a background gene regulatory network (GRN) is build using the network inference package CoRegNet [2]. This GRN allows to calculate the regulators influence, a sample-specific statistical value corresponding to an estimation of the transcription factors (TF) activities. By integrating the reverse engineered gene regulatory network into the metabolic model (CoRegFlux [5]) and learning from the regulators influence, our model can predict the metabolic genes expression levels in context-specific conditions. These predicted expressions are then converted into constraint for flux balance analysis leading to phenotype prediction and possible calculation of biomass-product coupled yield. Using data from *S. cerevisiae*, we applied our method to a high-dimensional gene expression dataset to infer a background gene regulatory network and compared the resulting phenotype simulations with those obtained by other relevant methods. Our method was shown to have a better performance and robustness to noise and was successfully used to study complex context-specific phenotype such as diauxic shift [5]:

More specifically, CoRegCAD aims at providing a set of functions to simulate the engineering of the regulatory network as well as relevant gene knock-out or over-expression. These simulations will then be used to optimize the best constructions to improve production and to select the most appropriate regulatory element to be included in the expression cassette in the chassis organism. The determination of its optimal insertion point within the genome to maximize the clustering of co-regulated genes will also be considered (GREAT [3])

Case-study on an industrial chassis : *Y. lipolytica*

To demonstrate the relevance of our strategy for less common organism of industrial interest, these methods will be developed and tested in *Y. lipolytica*, an oleaginous yeast whose metabolism is prone to lipid accumulation under conditions of nitrogen limitation. Following the CoRegCAD framework, a regulatory network consisting of 111 TF, 4451 target genes and 17048 regulatory interactions (YL-GRN-1) was inferred. Interrogation of this network highlighted the relevance of our method to identify several regulatory state corresponding to the yeast adaptation to nitrogen depletion. Using influence, we were also able to identify potential regulators and drivers of lipid accumulation, some of which were tested in the lab with 6 out of 9 being validated for their impact on lipid accumulation [6].

This work will provide proof-of-concept for the context-specific design of metabolic pathways of interest, by improving the yield under specific constraints.

Conclusions

While further development still need to be carry out, CoRegCAD purpose is to provide a framework relying on network inference and interrogation to guide the metabolic engineering of industrial chassis and achieve higher production of metabolite of interest in context-specific conditions. Using CoRegCAD, researchers will be able to reduce time-consuming and costly laboratory effort, to carry out functionalities studies and to identify regulatory element of interest for context-specific expression through the interrogation step and iterative learning process.

E. Appleton, C. Madsen, N. Roehner, and D. Densmore, "Design automation in synthetic biology," *Cold Spring Harb. Perspect. Biol.*, vol. 9, no. 4, 2017.

R. Nicolle, F. Radvanyi, and M. Elati, "CoRegNet: reconstruction and integrated analysis of co-regulatory networks.," *Bioinformatics*, vol. 31, no. 18, pp. 3066–8, 2015.

C. Bouyioukos and M. Elati, "GREAT: a web portal for Genome Regulatory," vol. 44, no. May, pp. 77–82, 2016.

E. Celińska, R. Ledesma-Amaro, M. Larroude, T. Rossignol, C. Pauthenier, and J. M. Nicaud, "Golden Gate Assembly system dedicated to complex pathway manipulation in Yarrowia lipolytica," *Microb. Biotechnol.*, vol. 10, no. 2, pp. 450–455, 2017.

D. T. Banos, P. Trébulle, and M. Elati, "Integrating transcriptional activity in genome-scale models of metabolism," *BMC Syst. Biol.*, vol. 11, no. Suppl 7, 2017.

P. Trébulle, J.-M. Nicaud, C. Leplat, and M. Elati, "Inference and interrogation of a coregulatory network in the context of lipid accumulation in Yarrowia lipolytica," *npj Syst. Biol. Appl.*, vol. 3, no. 1, p. 21, 2017.

Keywords: regulatory network, genome, scale modeling, computer, aided design, metabolic engineering

Integrative analysis of genomic regulation combining cistrome, epigenome and transcriptome

Lucie Khamvongsa-Charbonnier * [1]

[1] Theories and Approaches of Genomic Complexity (TAGC) – Aix Marseille Université, Institut National de la Santé et de la Recherche Médicale : U1090, Centre National de la Recherche Scientifique – Théories et approches de la complexité génomiqueParc scientifique de Luminy163 avenue de Luminy13288 Marseille cedex 9, France

Integrative analysis of genomic regulation combining cistrome, epigenome and transcriptome
Lucie Khamvongsa-Charbonnier1,2

Supervisors: Jacques van Helden1 and Ghislain Bidaut 2

1. Aix Marseille Univ, INSERM, TAGC, Marseille, France

2. Aix Marseille Univ, CNRS, INSERM, Institut Paoli-Calmettes, CRCM, Marseille, France

Gene expression is regulated by multiple mechanisms, which include the specific action of transcription factors and chromatin modulators. Transcription factors (TFs) bind DNA and interact with the transcriptional machinery. A variety of NGS technologies provide a genome-scale description of different mechanisms of regulation (transcription factor binding locations, histone modifications, DNA accessibility, DNA methylation, wide-range interactions between chromosome regions, ...) and of their impact on gene transcription.

RNA-seq and ChIP-seq are respectively used to measure gene expression and to obtain genome-wide maps of transcription factor occupancies and epigenetic signatures. Integrating these data types can help elucidating gene regulatory mechanisms. However, their exploitation has not yet reached its full potential, due to a limited availability of multi-omics data integration tools and methods.

The goal of my thesis is to develop, assess and apply bioinformatics approaches to integrate ChIP-seq and RNA-seq data, that extend beyond the usual intersection between list of genes (differentially expressed, associated to TF binding regions or chromatin marks), and keep track of the quantitative nature of the data until the integration stage.

We are currently focusing on canonical correlations analysis (CCA) and its generalized form (GCCA), which are multivariate statistical methods allowing the identification of relationships

[*] Speaker

(correlations) between two or more tables of quantitative variables. We will apply these methods to decipher the combinatorial relationships between transcription factors, epigenomics marks and transcription.

Among the questions to be addressed, we will evaluate different indicators of the relationship between cis-regulatory regions and target genes (number of peaks, read density, intronic/upstream/downstream location, distance to the TSS, ...), build combinatorial models and assess their accuracy as input variables to predict condition-dependent transcription levels.

These approaches will be tested on 2 or 3 study cases for which we dispose of detailed information on chromatin factors and/or marks and transcriptional response. The relevance of the results will be evaluated in direct collaboration with the biologists who produce this data.

The concepts and approaches will be illustrated based on concrete results on the impact of Hox mutations on Drosophila transcriptome, led in collaboration with Andrew Saurin (IBDM, Marseille).

Keywords: NGS, Cistrome, Epigenome, Transcriptome, regulation, ChIp, seq, RNA, seq, Integration

Building artificial genetical genomic datasets to optimize the choice of gene regulatory network inference methods.

Lise Pomiès * [1], Louise Gody [2], Charlottte Penouilh-Suzette [2], Nicolas Langlade [2], Brigitte Mangin [2], Simon De Givry† [1]

[1] Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT INRA) – Institut National de la Recherche Agronomique : UR875 – Chemin de Borde Rouge, 31320 Castanet Tolosan, France
[2] Laboratoire des interactions plantes micro-organismes (LIPM) – Institut National de la Recherche Agronomique, Centre National de la Recherche Scientifique : UMR2594 – Chemin de Borde-Rouge - BP 27 31326 CASTANET TOLOSAN CEDEX, France

One of the central targets of Systems Biology is to decipher the complex behavior of a living cell in its environment. A gene regulatory network is a simplified representation of the gene-level interactions. Network inference methods are powerful tools to understand such complex biological processes [1]. However, it could be difficult to identify an algorithm adapted to a specific experimental dataset. Artificial datasets could be used to test different algorithms of inference and select the most accurate one. But, available artificial datasets are more like ideal datasets and consequently quite different from measured datasets. In our case, we didn't know if classical network inference algorithms (like bayesian network, mixed model, penalised regression, random forests, or a combination of them [2]) will work correctly on our dataset. We also didn't find an artificial dataset with the same properties as our experimental data. This is why, we decided to create our own artificial dataset. We present here the characteristics of our experimental dataset and the strategy we elaborate to create an artificial dataset with the same properties as our experimental dataset.
We work on domesticated sunflower (Helianthus annuus), a highly resistant crop plant to drought. The sequencing of the genome of the XRQ line of sunflower, had been published last year [3]. In the context of climate changes it's interesting to understand how sunflower resists to drought at the molecular level and how this resistant interacts with the phenomenon of heterosis when new varieties are created.

To answer this question, two transcriptomic experiments were performed. The goal of the first experiment was to select genes involved in response to drought and in the heterosis phenomenon. This experiment was performed on a hydric-control environment. Eight different parental genotypes of sunflower (4 males and 4 females) and their 16 hybrids were cultivated. The 8 parental genotypes are homozygous for all genes, their hybrids could be homozygous or heterozygous depending on the locus. Sunflowers were cultivated in two hydric conditions: (i) in drought condition and (ii) with sufficient water level. The expression levels of all genes were measured in both conditions and for all genotypes via RNA-sequencing. From those transcriptomics measurements we selected 180 genes responding to drought, heterosis and in interaction

*Speaker
†Corresponding author: simon.de-givry@inra.fr

between drought and heterosis. Because we are focusing on gene regulation we chose transcription factors (detected by iTAK [4] and plantTFCat [5]) to compose the main chunk of our dataset.

The goal of the second experiment was to collect enough data, to performed a gene regulatory network inference, on the 180 selected genes. The experiment was conducted on a field with 435 hybrids created from 72 homozygous parental genotypes (36 females and 36 males) including genotypes from the first experiment. The expressions of the 180 selected genes were measured by qPCR (Fluidigm technology) for all the hybrids. For all parental genotypes, single-nucleotide-polymorphisms (SNPs) were detected against the reference genome of sunflower (XRQ line).

The experimental dataset contains the expression of 180 genes on 435 hybrids. The collected data are not independent as they come from different parental genotypes and their hybrids. We don't know the effect of this dependency between the genotypes on network inference algorithms. To measure the impact of a non-independent dataset on existing network inference methods, we have created artificial datasets to test the accuracy of the methods.

First step is to create an artificial network. We decided to collect informations about interactions between our 180 genes of interest in different public databases. As the sequencing of the sunflower genome is recent, really few informations are available on databases for this plant. For this reason, we decided to collect interactions between the homologous genes of our selected genes on the plant model Arabidopsis thaliana. For 7 sunflower genes, no homologous genes were found. The homologous genes are the nodes of our artificial network. We collected interactions from 3 databases (i) AtPID, a database specific to A. thaliana containing interactions between proteins [6], (ii) AtRegNet specific to A. thaliana containing regulations between transcription factors and target genes [7], and (iii) PlantRegMap a plant database containing regulations between transcription factors and other genes [8]. The three databases contain links found in the literature, or resulting from experiments (as Chip-seq experiments). The third database also contains predicted regulations via detection of binding motifs, on the promoter of target genes, recognized by specific transcription factors. We selected in these databases only directed links, corresponding to expression regulations, involving two genes from our selection. We collected 364 regulations (36 in AtPID, 16 in AtRegNet, and 312 in PlantRegMap), 62% of these regulations were predicted regulations. Those 364 regulations form the edges of our artificial network. The type of regulation (activation or repression of the expression) is known for only two regulations. In the database AtRegNet where the nature of regulations are described, 64% were activation of the expression and 36% were repression of the expression. We decided to randomly associate each edge of our network to a particular type of regulation with a probability of 64% to be an activation of the expression, the rest being a repression of the expression.

In our experimental dataset, each parental genotype has a list of SNPs, associated to a score either 0 if it is like in XRQ-line or 1 if different. It is easy to deduce the SNPs of the hybrids by combining locus-per-locus the SNPs of their parents. In order to study the effect of genetic polymorphism on gene expressions, we created new virtual hybrid genotypes associated to DNA variants on each measured gene. To simplify this analysis, we considered one variant per gene. To be closer to our biological variety we created this DNA variants based on SNPs of the experimental data. For each gene of interest, we collected the SNPs present in their genomic sequence and their promoter region for each parental genotype. Using a K-medoid clustering with a Manhattan distance on the SNP data, genotypes were classified in two groups. The group of genotypes with SNPs close to the SNP values of XRQ-line has a DNA variant score of 0, and the other group of genotypes has a score of 1, for this gene. For hybrids, the score of DNA variant on each gene is equal to the mean score of their parents. It can take 3 values: 0 or 1 if the hybrid is homozygous for this gene, or 0.5 if it is heterozygous. We now had a collection of hybrids, with known DNA variations on our genes of interest.

The third step is to produce artificial measures of expression for the selected genes. The data simulator SysGenSIM simulates steady state gene expressions for different genotypes using ordinary differential equations [9]. The simulation is based on a gene network topology and DNA variant for each gene. In their model, each gene has only one DNA variant. The DNA variant of a gene has either a cis-effect (meaning it influences the rate of transcription of the gene) or a trans-effect (meaning it modifies the efficiency of the gene regulation activity). The equation describing the accumulation of a gene transcript for a given genotype is composed of two parts. The first part of the equation describes the rate of expression of the gene, and the second part describes the rate of degradation of the transcript. The expression rate is modulated by the effect of the DNA variant of the gene and the expression of the regulators of this gene in the network. The DNA variant of the regulators have also an impact on the efficiency of the regulation. For the moment, SysGenSIM only works on recombinant inbred lines (RIL). We slightly modified the simulator to use our heterogenous hybrids, and mimetized the allelic dominance caused by the heterosis phenomenon. In case of genes with heterozygous DNA variant, the DNA variant effect is randomly chosen, with a probability of 0.8 to be an additive effect of the DNA variant effect of both parents and a probability of 0.2 to be a dominant effect of the DNA variant effect of one parent. With the modified version of SysGenSIM we can produce artificial gene expression data for the artificial gene network and hybrids we previously generated.

To adjust the different parameters of SysGenSIM, to produce a dataset as close as possible to our real data, we estimated the part of the variance explained by the genotypes (also called heritability) in the produced dataset and in the real dataset. This heritability is calculated via a mixed model [10].

By choosing at random the type of regulation (activation or repression), the DNA effect (cis- or trans-effect), and the allelic dominance effect for heterosis we produced different simulated gene expression datasets for our 180 genes and 435 genotypes. For each dataset, a particular gene regulation network with the same topology is associated. As a consequence, we can now test different methods of network inference and test the accuracy of these methods by comparing networks produced by the algorithms to the true network. Network inference methods with the best results will be used on the experimental dataset to answer our biological question.

In conclusion, we have developed a strategy to create an artificial dataset of gene expression measurements. The aim of this dataset is to test and select network inference methods adapted to a non-independent dataset for understanding the response to drought and heterosis phenomenon of sunflowers. The strategy is constituted of the following 4 steps, that could be adapted for other biological experiments and other types of data :

(i) Construction of an artificial network based on real biological information available for the same biological process on a close organism ;

(ii) Creation of artificial hybrid genotypes based on genomic information available for the real hybrids used in the experiment ;

(iii) Selection and adaptation of a data simulator emulating the same type of experiment that the one we performed, with steady state measurements on different genotypes ;

(iv) Comparison of the biological score obtained on real and simulated datasets (in our case the heritability score) to adjust parameters of the simulator.

For each step it's important to use real biological information to in the end obtain an artificial dataset with biological properties close to properties of the real one. Doing like this, we hope the probability that networks inference methods perform the same in simulated as in real data is really high.

Banf & Rhee (2017). Computational inference of gene regulatory networks: Approaches, limitations and opportunities. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1860(1), 41–52.

Vignes et al. (2014). Gene Network Inference, chapter A Panel of Learning Methods for the Reconstruction of Gene Regulatory Networks in a Systems Genetics Contex. Springer.

Badouin et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature, 546(7656), 148–152.

Zheng et al. (2016). iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. Molecular Plant 9, 1667–1670

Dai et al. (2013). PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. BMC Bioinformatics, 14:321.

Li et al. (2011). AtPID: The overall hierarchical functional protein interaction network interface and analytic platform for arabidopsis. Nucleic Acids Research, 39(SUPPL. 1), 1130–1133.

Palaniswamy et al. (2006). AGRIS and AtRegNet. A Platform to Link cis-Regulatory Elements and Transcription Factors into Regulatory Networks. Plant Physiology, 140(3), 818–829.

Jin et al. (2017). PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Research, 45(D1), D1040–D1045.

Pinna et al. (2011). Simulating systems genetics data with SysGenSIM. Bioinformatics, 27(17), 2459–2462.

Mangin et al. (2017). Genomic Prediction of Sunflower Hybrids Oil Content. Frontiers in Plant Science, 8, 1–12.

Keywords: gene regulatory network, artificial dataset, sunflower

A Genetic Algorithm to identify biological processes deregulated in Progeria

Elva María Novoa Del Toro * [1], Efrén Mezura-Montes , Matthieu Vignes , Elisabeth Remy , Diane Frankel , Alexandre Atkinson , Annachiara De-Sandre-Giovannoli , Patrice Roll , Lévy Nicolas , Laurent Tichit [2], Anaïs Baudot [3]

[1] Aix Marseille Univ, INSERM, MMG, Marseille Medical Genetics, Marseille, France – Aix Marseille Univ, INSERM, MMG, Marseille Medical Genetics, Marseille, France – France
[2] AMU – Aix-Marseille Université - AMU – France
[3] Marseille medical genetics - Centre de génétique médicale de Marseille (MMG) – Aix Marseille Université : UMR$_S$1251, $InstitutNationaldelaSantéetdelaRechercheMédicale : UMR_S1251, CNRS : UMRS1251 - -FacultédeMédecine - Timone27, boulevardJeanMoulin13385Marseillecedex5, France$

Elva-María Novoa-del-Toro [1], Efrén Mezura-Montes [2], Matthieu Vignes [3], Elisabeth Remy [4], Diane Frankel [1], Alexandre Atkinson [1], Annachiara De-Sandre-Giovannoli [1],

Patrice Roll [1], Nicolas Lévy [1], Laurent Tichit [4], Ana´is Baudot [1]

Aix Marseille Univ, INSERM, MMG, Marseille Medical Genetics, Marseille, France

University of Veracruz, Artificial Intelligence Research Center, Mexico

Massey University, Institute of Fundamental Sciences, New Zealand

Aix Marseille Univ, CNRS, I2M, Marseille Institute of Mathematics, Marseille, France

The progeria syndrome is a very rare genetic disease characterized by a premature and accelerated aging. Progeria is caused by mutations in the *LMNA* gene that lead to the production of a truncated and toxic protein called progerin. Accumulation or this protein in nuclei induces nuclear shape deformations associated with dysfunctions of many cellular processes, such as gene expression, leading to a premature senescence.

To better understand the cellular consequences of *LMNA* mutations, we conducted a mRNA/miRNA sequencing of five progeria and three control fibroblast samples. We further developed a network-based approach in order to improve the functional interpretation of the observed differentially expressed mRNA and miRNAs. Our objective is to identify the cellular functions deregulated in progeria, which is equivalent to find subnetworks both highly interconnected and enriched in up- or down-regulated mRNA/miRNA.

*Speaker

To this goal, we designed a Genetic Algorithm (GA). We built first a network merging protein-protein and miRNA-mRNA interactions. In this network, every node (representing either a miRNA or a protein/mRNA) is associated to a Z-score, reflecting its differential expression in patient versus control samples. The GA generates a population of individuals, where each individual is a potential solution to the problem, in our case, a subnetwork. Every individual is then evaluated by a fitness function, which accounts for the clustering coefficient - as a local connectivity degree, and the average normalized Z-score - as a measure of local gene activity. At each generation, the fittest individuals of the population have a higher probability to mate, and hence to transmit their features to the next generation. The GA stops after a fixed number of generations, and returns the individuals maximizing the above-mentioned fitness function.

Each individual is a connected subnetwork enriched in differentially expressed mRNA and miRNA, but that can also contain non-differentially expressed nodes if they link differentially expressed ones. Our approach thereby allows us integrating transcriptomics datasets, enhancing the detection of relevant mRNA and miRNA, and reveals cellular processes deregulated in progeria.

It is to note that our approach is generic, and can be applied to any transcriptomics datasets. In the future, we will extend the GA to leverage multiplex networks, i.e., networks composed of multiple layers of physical or functional interactions, such as PPI, molecular complexes and co-expression.

Keywords: Genetic algorithm, network, progeria

Studying tomato fruit development through reconstruction of stage specific genome scale metabolic models using proteomics and transcriptomics

Sylvain Prigent *† 1

[1] Biologie du fruit et pathologie (BFP) – Institut National de la Recherche Agronomique : UMR1332 – Centre INRA Bordeaux-Aquitaine 71 avenue Bourlaux BP81 F-33883 Villenave d'Ornon, France

Tomato fruit development is a highly complex process involving several steps. It has been intensively studied, making it an important model for fruit development. While being a continuous process, it is usually divided into several stages corresponding to phenotypic stages of the fruit: immature green, mature green, orange-breaker and red ripening. During this development, metabolism of the tomato fruit is highly re-organized to adapt to any special needs of fruits during the different periods, like starch degradation at the end of the expansion phase, or a big need of soluble sugars in the vacuole at the end of the maturation.

Small scale metabolic models, both stoichiometric and kinetic based have already been reconstructed. Those models have been constrained using quantitative metabolomic data as well as enzyme activity measurements. While they have already proven useful to study the primary carbon metabolism through the entire developmental process of the fruit, those small scale data do not really enable a comprehensive study of the metabolism, especially when secondary metabolism and production of pigments become important.

Nowadays, several large scale data on tomato fruits exist, at different biological levels, including genomics, transcriptomics and proteomics. Based on quantitative RNA-seq and proteomics, we have developed a mathematical model enabling an estimation of synthesis and degradation rate constants indicating that, in tomato fruits, synthesizing a protein take 2 minutes 30 seconds, while its lifetime is around 11 days (median values). Integrating the information of presence/absence of enzymes, as a function of time, in metabolic models will enable us to reconstruct stage specific genome scale metabolic models. This will enable us to study those models both from a topological and a flux based point of views, to study metabolism all along the development of the fruit. Those models will also be useful to have a better understanding of metabolic diseases as well as colonization by fungus like mildew.

Keywords: Metabolism, Metabolic networks, proteomics, data integration

*Speaker

†Corresponding author: sylvain.prigent@inra.fr

IDENTIFICATION DES FACTEURS DE L'HÔTE FYN et PTPN11 IMPLIQUÉS DANS L'INFECTION PAR LE VIRUS DE L'HÉPATITE B.

Mohcine Elmessaoudi-Idrissi * [1,2], Anass Kettani [2], Pascal Pineau [3], Soumaya Benjelloun [1], Sayeh Ezzikouri [1]

[1] Unité de virologie, Laboratoire des Hépatites Virales, Institut Pasteur du Maroc – Morocco
[2] Laboratoire Biologie et Santé – URAC34, Equipe Modélisation Moléculaire et Contrôle Qualité, Faculté des Sciences Ben Msik, Université Hassan II de Casablanca – Morocco
[3] Unité Organisation Nucléaire et Oncogenèse , INSERM U993, Institut Pasteur, Paris – Institut Pasteur de Paris – France

Contexte : L'infection par le virus de l'hépatite B (VHB) représente un problème de santé publique majeur. Les
facteurs de l'hôte impliqués dans l'infection par le VHB sont complexes, hétérogènes et peu connus.
Objectif : Dans le but d'identifier des signatures d'expression génique et les processus biologiques impliqués, nous avons mené une
méta-analyse intégrée de données de microarray de patients de tissus hépatiques infectés par le VHB.
Matériel et méthodes : Trois jeux de données microarrays (142 patients infectés par le VHB et 19 témoins négatifs) ont été collectés.
Les données sont analysées à l'aide de l'outil bioinformatique "Networkanalyst" en se basant sur les approches combinées d' Effect
Size (ES).
Résultats : Au total, nous avons identifié 243 gènes différentiellement exprimés . L'enrichissement fonctionnel et l'analyse des
pathways ont révélé les voies du métabolisme des pyrimidines et de l'histidine, de dégradation de l'ARN et de la p53 comme termes
les plus enrichis. Dans l'analyse du Gene Ontology, la cytokinèse est le processus biologique le plus significativement enrichi.
Enfin,les réseaux d'interaction ont identifié FYN et PTPN11, deux gènes modulant l'immunité innée, comme les plus connectés avec
d'autres gènes.
Conclusion : Les résultats de la méta-analyse des données transcriptomiques ont montré que plusieurs mécanismes complexes et
hétérogènes sont impliqués dans l'infection par le VHB.

*Speaker

Keywords: Hépatite B, Bioinformatique, Microarrays, Méta, analyse, pathways, cible thérapeutique

Analyse du génome de tumeurs neuroendocrines par puces SNP

Mario Neou * [1], Chiara Villa † [1], Karine Perlemoine‡ [1], Victoria Verjus Lisfranc§ [1], Jouinot Anne¶ [1], Baussart Bertrand [2], Jerome Bertherat [1], Gaillard Stéphan [2], Guillaume Assié [1]

[1] Institut Cochin (UM3 (UMR 8104 / U1016)) – Université Paris Descartes - Paris 5 : UM3, Institut National de la Santé et de la Recherche Médicale : U1016, Centre National de la Recherche Scientifique : UMR8104 – 26 rue du faubourg saint jacques , 75014 Paris, France
[2] Hôpital Foch [Suresnes] – Hôpital Foch [Suresnes] – 40 Rue Worth 92150 Suresnes, France

Les tumeurs endocrines ont un comportement variable. Nous supposons que cette variabilité phénotypiques est en rapport avec une variabilité moléculaire à l'échelle cellulaire. Cette variabilité inclut notamment des anomalies chromosomiques spécifiques de chaque tumeur.

Les anomalies chromosomiques de 96 tumeurs endocrines ont été analysées par puce SNP. Ce travail rapporte une méthode originale d'analyse, et les anomalies chromosomiques identifiées.

Méthode :
Les patients inclus ont signé un consentement écrit, validé par le comité d'éthique local. L'ADN tumoral a été extrait et hybridé à des puces Illumina InfiniumCore-24v1-1 ; 300k SNPs.
Le calling des segments a été fait avec GAP (Popova et al, 2009) . Les données ont été ensuite étudiés grâce à des scripts originaux écrits en R.

Résultats :
L'analyse par GAP a identifié 60391 segments sur 96 tumeurs, réduits à 4611 après lissage et validation manuelle. Chaque segment est décrit en termes de nombre de copies absolues d'ADN, de ratio allélique, de pourcentage de cellule avec l'anomalie (sous-clonalité). En outre la plo´idie des tumeurs a été déterminé.
Nous avons également implémenté une méthode originale de calling de l'homozygotie germinale, correspondant à des régions en UPD (Uniparental Disomy), combinant le nombre de SNP et leur hétérozygotie.
Les régions les plus souvent gagnées sont sur les bras 7p, 7q, 19p... (30 %, 20 % et 20% respectivement).
Les régions les plus souvent perdues sont sur les bras 1p,23p ,2q... (13 %, 13 % et 10% respectivement).
15 délétions homozygotes et 7 amplicons (> 5 copies) ont été répertoriées.
124 régions avec une homozygotie germinale importante ont été identifiées, particulièrement

*Speaker
†Corresponding author: cm.villa@hopital-foch.org
‡Corresponding author: karine.perlemoine@inserm.fr
§Corresponding author: victoria.verjus@yahoo.fr
¶Corresponding author: anne.jouinot@aphp.fr

concentrés dans 5 échantillons.

Conclusion :
Cette analyse exploite toute l’information des puces à la recherche des anomalies chromosomiques associées à ce type tumoral mal caractérisé à ce jour.

Keywords: SNP Arrays, Chromosomal abberations

Integrating, visualising and jointly exploiting heteroclite knowledge to analyze omics data with NeOmics

Ludovic Léauté [1], Paul Dubos [1], Patricia Thébault* [2], Raluca Uricaru †‡ [1]

[1] Laboratoire Bordelais de Recherche en Informatique (LaBRI) – Université de Bordeaux, CNRS UMR 5800, Université de Bordeaux – Domaine Universitaire 351, cours de la Libération 33405 Talence Cedex, France

[2] Laboratoire Bordelais de Recherche en Informatique (LaBRI) – CNRS : UMR5800, Université de Bordeaux (Bordeaux, France), CNRS UMR 5800 – Domaine Universitaire 351, cours de la Libération 33405 Talence Cedex, France

Integrating, visualising and jointly exploiting heteroclite knowledge to analyze omics data with NeOmics

Introduction

Recent technological revolutions strongly impacted the omics data production for the last decades. The resulting data deluge is re-designing research in Biology by opening new opportunities in trans-omics (Yugi et al., 2016) that aim at capturing several levels of organization and/or information in the cell. The main objectives of these analysis is related to the understanding of genotype-phenotype relationships by identifying groups of differentially expressed genes according to different experimental conditions and/or tissues. In regards to the features of these conditions, prioritizing genes is essential for explaining a phenotype of interest. In this context the integrative analysis of omics data is becoming the new bottleneck in bioinformatics research. Several issues have to be considered, ranging from the diversity of omics data to the diversity of analysis approaches. For example, as defined by (Yugi et al., 2016), horizontal integrative meta-analysis rely on the integration of the same type of analysis that have been carried out with multi experimental conditions, whereas vertical integrative meta-analysis define an integration of multi types of omics data.

Firstly, the integration of multiple types of omics data is crucial to improve our understanding of the complexity of the cell life. Indeed, overlaying different points of view given by transcriptomics, proteomics and metabolomics analysis, allows designing new predictive methods.

[*] Corresponding author: patricia.thebault@labri.fr

[†] Speaker

[‡] Corresponding author: raluca.uricaru@u-bordeaux.fr

Secondly, in order to analyze and compare high-throughput data, various bioinformatics, biostatistics and computer science approaches are of interest, each of them aiming to identify groups of genes that are differentially expressed under different experimental conditions, as well as groups of co-regulated genes (Gadgil et al., 2008). As the choice of a method is a difficult task, each approach having its strengths and weaknesses, the integration of several methods is more than relevant. It allows to compute overlap sets between results thus improving the reliability of the predictions, but also to capitalize on the richness of each method by putting together results coming from different sources (Simon et al. 2003, Valls et al., 2008) .

At last, the interactive interpretation of results, which is done by placing the expert in the center of the data analysis process, is a real added value. Guiding the analysis based on the effect of previous choices as well as on the interpretation of the data issued from the public databases, helps to improve the comprehension of the methods, as well as the specificities of the data and therefore encourages the generation of new hypotheses in the analysis process (Thébault et al., 2015).

Methods

In an ideal world, one would like to conglomerate numerous and highly heteroclite types of information corresponding to : biological data (genes, proteins, ARNs, ...), biological experiments (microarray analysis, RNA-Seq analysis, ...), analysis methods (issued from statistics, computer science, ...), results issued from the different analysis being performed. These need to be cross-referenced with diverse types of annotation information like Gene Ontology terms, biological pathways (metabolic, genetic, ...), disease related information, etc. The integration of all these types of information in a unique knowledge repository requires to reconsider the entire structuring of the information and the definition of graphical representations allowing to exploit at maximum and in an efficient way this highly complex repository.

Here we propose a method implemented in a prototype called NeOmics, meant for building and querying this kind of repository, and we show an example of analysis that can be carried out with our visualisation system. Specifically, the different pieces of information are modeled with a so-called "property graph", physically represented in a graph database with the Neo4j management system. With respect to relational databases, graph databases are the most adequate choice in our case, as they can cope with vast, highly heteroclite information and with evolutive contents. Compared to classical graph-like representations, graph databases have the advantages of offering a support for semantics, querying and of being able to manipulate datasets too large to be visualised (Lysenko I, 2016). Moreover, in order to capitalize on its bioinformatics specific functionalities, we chose to plug our Neo4j knowledge repository into Cytoscape, the state of the art biological network visualisation system.

Figure 1 : A schematic representation of the multi-layers knowledge repository modeled with NeOmics system.

More precisely, our knowledge repository can be organised in four main "layers", each layer being composed of possibly numerous nodes, as depicted in Figure 1 : (1) Experiments - modeling the experiment information, (2) Analysis - the different analysis being conducted on the data produced by the experiments, (3) Results - the results in terms of groups of genes, proteins etc. produced by each analysis and (4) Databases - the public databases providing annotation information. As suggested in Figure 1, nodes inside and between layers can be connected, eventually with multi-edges. The objective of such modeling could be, for example, to identify modules or groups of biological objects (e.g., genes), which are specific to a tissue and / or phenotype studied in a given experiment, by jointly exploiting results issued from different analysis methods and enrich them with annotation information; this can be pushed further by mutualising results obtained from different studies (experiments).

The richness of our model and of our visualisation system comes from the complete freedom the user has to adapt the model to his specific application. Moreover, the architecture of the graph database is not stuck in time, indeed the model is evolutive, thus allowing complete refactoring during the advancement of the project.

Results

Figure 2 : Screenshot of a subgraph visualised with NeOmics system through Cytoscape network visualisation tool. Two different levels of information are represented in this composite sub-graph: the genes (blue circles) and the results of three statistical methods that aim to infer differentially expressed gene sets, while differentiating up and down regulations (resp. green and red circles).

A first case application of the pipeline we introduce here as a proof of concept, was developed by making use of microarray data dedicated to the analysis of hormone signaling pathways in Arabidopsis thaliana (Vert al., 2005). The most interesting feature of these data relates to the need to consider an horizontal integration to conjointly use expression information coming from different microarray studies.

The different experiments and the entire set of analysis results were first loaded in our NeOmics repository, then queried with Cypher query language, finally the resulting subgraph is visualised with Cytoscape. As depicted in Figure 2, the model is centered around the data nodes, in our case the genes (blue circles). Additional level information involve the prior use of three state-of-the-art statistics methods for differential expression analysis : limma, rank product test and weighted average difference test (purple circles) that aim to infer gene modules. Then, the Neo4j's Graph Query Language (Cypher) is used to query the repository, for instance, to select the genes that are predicted as differentially expressed by several methods, or as shown in Figure 2 (B), the genes that meet at least one predictive method among the three. In general, a query should be driven by the main scientific questions that motivated the experimental data production. As the biological interpretation is crucial to exploiting any query result, an additional level of knowledge can be added by incorporating information from public databases like Gene Ontology, KEGG, etc. In this way, thanks to the interactive facilities, the biologist expert has the opportunity to explore, combine and interpret any type of pre-computed analysis.

Conclusion

Facing both the vast variety of omics data, the wide range of computational methods, and the multitude of biological knowledge, NeOmics is a first prototype to conglomerate numerous and highly heteroclite types of information. These different levels of knowledge can be ranged from experimental data to predictive results, while integrating biological information for interactive interpretation. The modular modelization of NeOmics is based on a multi-graph that offers to the user the possibility to adapt the model to his specific application. Moreover, the architecture of the graph database is not stuck in time, as the model is evolutive, thus allowing complete refactoring during the advancement of the project.

Bibliography

Gadgil M. A Population Proportion approach for ranking differentially expressed genes. BMC Bioinformatics. 2008; 9:380.

Lysenko A, Roznovăţ IA, Saqi M, Mazein A, Rawlings CJ, Auffray C. Representing and querying disease networks using graph databases. BioData Mining. 2016; 9:23.

Simon R, Radmacher MD, Dobbin K and McShane LM. Pitfalls in the use of DNA microarray data for diagnosis and prognostic classification. J Natl Cancer Inst. 2003; 95,14-18.

Thébault P, Bourqui R, Benchimol W, Gaspin C, Sirand-Pugnet P, Uricaru R, Dutour I. Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks. Brief Bioinform. 2015 ;16(5):795-805.

Valls J, Grau M, Solé X, Hernández P, Montaner D, Dopazo J, Peinado MA, Capellá G, Moreno V, Pujana MA. CLEAR-test: combining inference for differential expression and variability in microarray data analysis. J Biomed Inform. 2008 ;41(1):33-45.

Vert G, Nemhauser JL, Geldner N, Hong F, Chory J. Molecular mechanisms of steroid hormone signaling in plants. Annu Rev Cell Dev Biol. 2005;21:177-201.
Yugi K, Kubota H, Hatano A, et Kuroda S. Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers. Trends Biotechnol. 2016; 34 (4): 276-290.

Keywords: graph database, integrative omics analysis, visualization

Bacterial genome mining for the discovery of new natural products

Faustine Durand *† [1], Guillaume Letellier [1], Marine Rohmer [1], Georges Gaudriault [1]

[1] Deinove – Deinove – France

The intense use of antibiotics in the agribusiness field, combined with the fast evolution of bacterial populations, has caused the appearance of antibiotic resistant bacterial strains. Today, the emergence of resistant infectious diseases has become a major concern that may lead to 10 million deaths a year by 2050. Few novel antibiotic compounds have been discovered since two to three decades.

Deinove is a biotechnology company specialized in the development and production of compounds from natural origin using its library of 6,000 rare and extremophile bacteria. To answer the anti-bacterial resistance issue, Deinove's research platforms are focusing on the discovery of innovative natural products with anti-bacterial properties.

Indeed, with a few exceptions, most antibiotics are derivatives of secondary metabolites that are naturally synthesized by microorganisms. These are classified according to their biosynthetic pathways. There are 3 main families of secondary metabolites. On one hand, Non Ribosomal Peptides (NRPs) and Polyketides (PKs) are both synthesized by large modular enzymes. On the other hand, RiPPs (Ribosomally synthesized and post-translationally modified peptides) are synthesized ribosomally. On microorganisms' genomes the genes responsible for the biosynthesis of secondary metabolites are grouped together, forming Biosynthetic Gene Clusters (BGC). Bioinformatics methods exists, both to infer BGCs by mining the genome and also to predict the structure of the compound that could be biosynthesized. With the availability of low-cost sequencing and the large amount of already sequenced genomes, these methods can help drug discovery.

In this poster, examples of how different techniques such as genome mining and comparative genomics are used and combined for this purpose will be presented.

Keywords: Antibiotics, drug discovery, genome mining, screening platform, bioinformatics, bacteria

*Speaker

†Corresponding author: faustine.durand@etu.umontpellier.fr

Development of an automated pipeline to translate oligosaccharide sequences of glycosaminoglycans binding to proteins into 3D models.

Olivier Clerc * [1], Julien Mariethoz [2], Alain Rivet [3], Frederique Lisacek [2], Serge Pérez [3], Sylvie Ricard-Blum [1]

[1] Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) – CNRS : UMR5246, Université Claude Bernard - Lyon I (UCBL), Institut National des Sciences Appliquées [INSA] - Lyon, École Supérieure Chimie Physique Électronique de Lyon – Bâtiment CPE 43, Bld du 11 novembre 1918 69622 VILLEURBANNE CEDEX, France

[2] Proteome informatics Group, Swiss Institute of bioinformatics (SIB) – Geneva, Switzerland

[3] Centre de recherches sur les macromolécules végétales (CERMAV) – Université Joseph Fourier - Grenoble 1, Centre National de la Recherche Scientifique : UPR5301 – 601 Rue de la Chimie 38400 ST MARTIN D HERES, France

Mammalian glycosaminoglycans (GAGs) are linear complex polysaccharides divided into six groups: heparan sulfate (HS), heparin (HP), dermatan sulfate (DS), chondroitin sulfate (CS), keratan sulfate (KS) and hyaluronan (HA). They consist of disaccharide repeating units (a hexuronic acid or a hexose and a hexosamine). They undergo several modifications (e.g. epimerization, N- and O-sulfation and N-acetylation) during their biosynthesis, which provide them with a huge structural diversity. There are 69 unique disaccharides of GAGs and 23 HS disaccharides have been identified in mammals. GAGs bind to numerous extracellular, membrane and cellular proteins (more than 600 proteins bind to HP/HS), and these interactions mediate their biological activities. The structural and functional characterization of GAG-protein interactions is thus required to decipher their molecular mechanisms of action within the extracellular matrix (ECM) and at the cell surface. We have created an interaction database, MatrixDB (http://matrixdb.univ-lyon1.fr/, Chautard *et al.* 2011, Launay *et al.* 2015), focused on the ECM, to store protein-protein and protein-GAG interaction data. These data are manually extracted from the literature following the curation rules of the International Exchange Consortium (IMEx, https://www.imexconsortium.org/, Orchard *et al.*, 2012) *via* the curation interface of the IntAct database (https://www.ebi.ac.uk/intact/, Orchard *et al.*, 2014). However, a standard nomenclature and a machine-readable format of GAGs together with bioinformatics tools for mining GAG interaction data are lacking (Ricard-Blum, 2017, Ricard-Blum and Lisacek, 2017). We report here the building of an automated pipeline *i*) to standardize GAG sequences interacting with proteins manually curated from the literature, *ii*) to translate them into the machine-readable GlycoCT format (Herget *et al.*, 2008) and into SNFG (Symbol Nomenclature For Glycan) images (Varki *et al.*, 2015) and *iii*) to build their 3D models based on conformational maps. These maps were validated by the conformations of GAGs found in crystallized GAG-protein complexes, and were used to classify GAGs into eight major families.

*Speaker

Our pipeline comprises several steps and uses different bioinformatic tools. The first step is the manual conversion of the GAG sequences extracted from the literature, which are in various formats, into the IUPAC-condensed format used by the ChEBI database (Hasting*s et al.,* 2016). This format is then converted into the GlycoCT format with the SugarConverter (https://bitbucket.org/sib-pig/sugar-converter/) tool we developed. These formats (IUPAC and GlycoCT) are now used for cross-referencing MatrixDB, ChEBI and the GlyTouCan repository (Tiemeyer *et al.,* 2017). The third step is the translation of the GlycoCT format into SNFG images with GlycanBuilder (Ceroni *et al.*, 2007) and the GlycoWorkBench library (Damerell *et al.*, 2012). Both SugarConverter and GlycanBuilder can be used in command line to process a high number of GAG sequences and the SNFG nomenclature has been extended to include the 1C4 and 2S0 conformations of iduronic acid. Then the GlycoCT format is automatically converted *via* a converter we have developed (https://github.com/OlivierClerc/convert-glycoct-inp) into the INP format required by an open source software package building 3-dimensional structures of polysaccharides (POLYS 2.0, Engelsen *et al.*, 2014). POLYS is used to generate 3D models of GAG sequences from conformational maps [(ϕ, ψ) potential-energy surface] of the 30 most frequent disaccharides occurring *in vivo* and in crystallized protein-GAG complexes. The conformational maps were validated with a dataset of 74 crystallized GAG-protein complexes, retrieved from the Protein Data Bank, through the calculation of phi and psi angles between monosaccharides with the CARP tool (CArbohydrate Ramachandran Plot, http://glycosciences.de/tools/carp, L'utteke *et al.*, 2005). The 3D models of GAGs generated as PDB files, are integrated into MatrixDB database and visualized on the Biomolecule Report page with LiteMol, a tool handling 3D macromolecular data developed by D. Sehnal's group (http://webchemdev.ncbr.muni.cz/Litemol/, Sehnal *et al.*, 2017). These models and experimental 3D structures of GAGs (when available) will be integrated into the GAG-protein interaction network built from MatrixDB data to cluster proteins binding to the same 3D GAG structures.

References

Ceroni A, Dell A, Haslam SM. The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. Source Code Biol Med. 2007 Aug 7;2:3.

Chautard E, Fatoux-Ardore M, Ballut L, Thierry-Mieg N, Ricard-Blum S. MatrixDB, the extracellular matrix interaction database. Nucleic Acids Res. 2011 39: D235-40.

Damerell D, Ceroni A, Maass K, Ranzinger R, Dell A, Haslam SM. The GlycanBuilder and GlycoWorkbench glycoinformatics tools: updates and new developments. Biol Chem. 2012 393:1357-62.

Engelsen SB, Hansen PI, Pérez S. POLYS 2.0: An open source software package for building three-dimensional structures of polysaccharides. Biopolymers. 2014 101:733-43.

Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. 2016 44:D1214-9.

Herget S, Ranzinger R, Maass K, Lieth CW. GlycoCT-a unifying sequence format for carbohydrates. Carbohydr Res. 2008 343:2162-71.

Launay G, Salza R, Multedo D, Thierry-Mieg N, Ricard-Blum S. MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. Nucleic Acids Res. 2015 43: D321-7

L´utteke T, Frank M, von der Lieth CW. Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. Nucleic Acids Res. 2005 33:D242-6.

Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FS, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock RE, Hannick LI, Jurisica I, Khadake J, Lynn DJ, Mahadevan U, Perfetto L, Raghunath A, Ricard-Blum S, Roechert B, Salwinski L, St´umpflen V, Tyers M, Uetz P, Xenarios I, Hermjakob H. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods. 2012 9:345-50.

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014 42:D358-63.

Ricard-Blum S. Protein–glycosaminoglycan interaction networks: Focus on heparan sulfate. Perspect Sci. 2017 11: 62-69.

Ricard-Blum S, Lisacek F. Glycosaminoglycanomics: where we are. Glycoconj J. 2017 34:339-349.

Sehnal D, Deshpande M, Vařeková RS, Mir S, Berka K, Midlik A, Pravda L, Velankar S and Koča J. LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. Nat. Methods, 2017 14, 1121–1122.

Tiemeyer M, Aoki K, Paulson J, Cummings RD, York WS, Karlsson NG, Lisacek F, Packer NH, Campbell MP, Aoki NP, Fujita A, Matsubara M, Shinmachi D, Tsuchiya S, Yamada I, Pierce M, Ranzinger R, Narimatsu H, Aoki-Kinoshita KF. GlyTouCan: an accessible glycan structure repository. Glycobiology. 2017 27:915-919.

Varki A, Cummings RD, Aebi M, Packer NH, Seeberger PH, Esko JD, Stanley P, Hart G, Darvill A, Kinoshita T, Prestegard JJ, Schnaar RL, Freeze HH, Marth JD, Bertozzi CR, Etzler ME, Frank M, Vliegenthart FG, L´utteke T, Perez S, Bolton E, Rudd P, Paulson J, Kanehisa M, Toukach P, Aoki-Kinoshita KF, Dell A, Narimatsu H, York W, Taniguchi N and Kornfeld S. Symbol Nomenclature for Graphical Representations of Glycans, *Glycobiology*, 2015 25, 1323-1324.

Keywords: Biomolecular interactions, Glycosaminoglycans, 3D models, Database

Measuring the expression of RNA sequences in hundreds of RNA-seq libraries at nucleotide resolution

Mariam Bouzid * , Jérôme Audoux [1], Rayan Chikhi [2], Daniel Gautheret [3]

[1] SeqOne – Institut National de la Santé et de la Recherche Médicale - INSERM – c/IRMB, Hôpital St Eloi, Montpellier., France

[2] CRIStAL, Université de Lille – Centre national de la recherche scientifique - CNRS (France) – Lille, France

[3] Institut de Biologie Intégrative de la Cellule (I2BC) – Université Paris-Sud - Paris 11, Commissariat à l'énergie atomique et aux énergies alternatives : DRF/I2BC, Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR9198 – Bâtiment 21, 1 avenue de la Terrasse, 91198 Gif/Yvette cedex, France

With NGS data repositories now breaking the million libraries and ten petabyte barrier, performing sequence search in this data requires ultra-efficient tools. Recent data structures such as Sequence Bloom Trees (SBT), Mantis and SeqOthello propose efficient indexing strategies, which have been successfully applied to perform sequence queries in up to 450,000 NGS libraries, returning results in seconds. However, a major limitation of these technologies is that they return only a binary, presence/absence information. This is not satisfying for gene expression studies or other analysis of RNA-seq data, since the important question in this field is how many times a given RNA sequence is observed in each dataset, which is used as a proxy for transcript abundance. The ability to measure the abundance of any arbitrary subsequence in large RNA-seq repositories would open the way to fascinating applications, such as profiling the expression of any specific transcript (for instance a novel splice variant or fusion RNA) in thousands of public libraries. Here we describe a comparative analysis of different indexing methods compatible with count retrieval. Our benchmark set is made of 550 prostate cancer RNA-seq libraries representing about 3 TB of compressed sequence data, or 55 billion sequence reads. Each sequence library is converted into a k-mer table and the resulting multiple-library k-mer index and count table are stored and queried using different strategies. These include (1) an implementation of the RocksDB key-value storage system, (2) a bgzip/Tabix combination, and (3) a minimal perfect hashing strategy. We report on the performances of each strategy in terms of index building speed, index size and query speed, varying different parameters such as the volume of sequence queries and RAM vs disk space of the indexes and count tables. Our results indicate that retrieving expression levels of exact arbitrary subsequences in such a large RNA-seq database is possible in the order of seconds, with a total index size reduced to about one tenth of the initial sequence file. This encouraging result opens the way to a new generation of expression analysis software that will enable profiling RNA variants in databases where such searches are currently impossible.

[*] Speaker

Keywords: RNA, seq, Gene expression, transcriptomics, data structure, NGS

Characterization of mouse early hematopoiesis changes during aging with high throughput single-cell RNA-sequencing

Léonard Hérault *† [1,2], Mathilde Poplineau [2,3], Nadine Platet [2], Elisabeth Remy‡ [1], Estelle Duprez§ [2]

[1] Institut de Mathématiques de Marseille (I2M) – Aix Marseille Université : UMR7373, Ecole Centrale de Marseille : UMR7373, Centre National de la Recherche Scientifique : UMR7373 – Centre de Mathématiques et Informatique (CMI)Technopôle Château-Gombert39, rue Frédéric Joliot Curie13453 Marseille Cedex 13, France

[2] Centre de Recherche en Cancérologie de Marseille (CRCM) – Aix Marseille Université : UM105, Institut Paoli-Calmettes : UMR7258, Institut National de la Santé et de la Recherche Médicale : U1068, Centre National de la Recherche Scientifique : UMR7258 – 27 bd Leï Roure, BP 30005913273 Marseille Cedex 09, France

[3] Department of Cellular and Molecular Medicine, Graduate School of Medicine, Chiba University (CellMolMed) – Chiba 260-8670, Japan

At the apex of hematopoiesis, a critical balance is maintained between self-renewal and lineage differentiation of long term hematopoietic stem cells (HSCs). With aging this balance is altered with an increase of self-renewal long term HSCs and a myeloid biased differentiation. This poster presents the implanted data analysis strategy and our first results that characterize the mouse HSC pool in bone marrow (BM) and its intrinsic cellular changes and variations during aging with single cell RNA-seq.
First, early mouse hematopoietic stem cell (Lineage-, SCA1+, KIT+, FLT3-) pool was purified by multi-parameter fluorescence-activated cell sorting (FACS) from bone marrow of young (3 month) and old (18 month) mice. Gene expression profiles of about 6000 young HSCs and 13000 old HSCs were obtained using the recent 10X Genomics single-cell 3'mRNA-seq technology.

Then, pseudotime-ordering was performed for each single cell data set, using Monocle [1]; an algorithm that orders cells along a differentiation trajectory potential with multiple branching, by operating reverse graph embedding. Unsupervised approach called dpFeatures [1] was applied to ordering the cells. The obtained differentiation trajectories were confirmed by visualizing expression evolution of two cellular surface markers: the CD34 and the CD48 that are differentially expressed along HSC differentiation. The estimation of cell cycle phase for each cell made with the cell-cycle predictor cyclone [2] highlighted the influence of cell cycle on single cell transcriptomic data, therefore we decided to use the cell cycle phase as a blocking factor in the analysis with Monocle.

Interestingly, taking into account cell cycle for pseudotime-ordering revealed a branching in

*Speaker

†Corresponding author: leonard.herault@inserm.fr

‡Corresponding author: elisabeth.remy@univ-amu.fr

§Corresponding author: estelle.duprez@inserm.fr early hematopoiesis with a split between cells that continue through ”classical differentiation” and cells that differentiate more quickly into an inflammatory state. This branching was observed in young as well as in old mice suggesting that HSC pseudotime-ordering was globally not modified with age.

However, comparison between young and old HSC pool highlights the known accumulation of long term HSCs (defined as CD34-, CD48-) in old mice. Furthermore, differential gene expression and enrichment tests pointed out several aging features previously observed at early states of hematopoiesis [3]. Indeed, our old HSC population showed an increase in oxidative phosphorylation in the mitochondrial electron transport chain, in hydrogen peroxide catabolism process and in translation rate, and a decrease in cellular response to DNA damage compared to young HSCs.

Conclusion: by applying an unsupervised ordering of single cell transcriptomic data with cell-cycle corrections, we were able to construct relevant differentiation trajectories for young and old HSCs. This data set is a starting point to build a model of HSC aging using new Boolean model inference from single-cell RNA seq methods [4].

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., & Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nature methods, 14(10), 979.

Scialdone, A., Natarajan, K. N., Saraiva, L. R., Proserpio, V., Teichmann, S. A., Stegle, O., ... & Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods, 85, 54-61.

Signer, R. A., & Morrison, S. J. (2013). Mechanisms that regulate stem cell aging and life span. Cell stem cell, 12(2), 152-165.

Fiers, M. W., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., & Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. Briefings in functional genomics.

Keywords: single cell RNA sequencing, hematopoietic stem cell, aging, pseudotime ordering, mouse hematopoiesis

Towards an integrated approach for the annotation of ICEs and IMEs in bacterial genomes

Julie Lao *† 1,2, Thomas Lacroix 1, Gérard Guédon 2, Nathalie Leblond-Bourget 2, Hélène Chiapello 1

[1] Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE) – Institut national de la recherche agronomique (INRA) : UR1404 – Bâtiment 210-233 Domaine de Vilvert 78350 Jouy en Josas Cedex, France

[2] Dynamique des Génomes et Adaptation Microbienne (DynAMic) – Université de Lorraine, Institut national de la recherche agronomique (INRA) : UMR1128 – Université de Lorraine, Faculté des Sciences et Technologies, Bd des Aiguillettes, BP 70239, 54506 Vandoeuvre-les-Nancy Cedex, France

Mobile genetic elements play a key role in bacterial genome evolution by enabling gene acquisition through horizontal gene transfer. Among these elements, some are integrated in the chromosome of their hosts and transferred by conjugation. There is two types of such elements: (i) Integrative Conjugative Elements (ICEs) which encode their own transfer and (ii) Integrative Mobilizable Elements (IMEs) which use the transfer machinery of co-resident conjugative element for their own transfer.
ICEs and IMEs are still poorly known and annotated in bacterial genomes. However early research suggests their high prevalence [1]. In addition to the genes necessary for their transfer or controlling their mobility, ICEs and IMEs carry cargo genes, such as antibiotic resistance genes, that can confer advantageous properties to the bacteria that carry them. Thus, they largely participate in the emergence of pathogenic multidrug resistant bacteria.

Different biological functions are needed for the transfer and maintenance of ICEs and IMEs into the recipient cell. Genes and sequences involved in the same biological function are grouped into modules. All ICEs have an integration module, a conjugation module, a regulation module and one or more adaptation modules. In IMEs, the mobilization module replaces the conjugation module and the regulation and adaptation modules may be absent.

Identification and precise delineation of ICEs and IMEs is a complex task that requires dedicated bioinformatics approaches. Two strategies currently exist:

(i) A pipeline that delineates ICEs in bacterial genomes by using the core genes that surrounds them [2,3]. This procedure is based on the detection of the conjugation module using the CONJscan module of the MacSyFinder software [4,5] and needs at least 4 different closely-related genomes to enable ICE annotation.

(ii) A semi-automated procedure that allows to delineate the boundaries of ICEs and IMEs

[*] Speaker

[†] Corresponding author: julie.lao@inra.fr

at the nucleotide level, and annotate their integration site in one genome. The procedure is based on the detection of signature proteins of the integration and conjugation modules extracted from known elements in streptococci [6]. This procedure makes it possible to annotate more complex structures of ICEs and IMEs, such as series of elements in accretion and elements inserted within another element.

Mobile genetic elements evolve rapidly mainly through acquisition, loss and exchange of modules thus making the detection and accurate annotation of ICEs and IMEs bounds a difficult task.

In this poster, we will present first results obtained from the comparison of the two strategies using a reference dataset containing known ICEs and IMEs in streptococci. The results of this benchmark suggest that it would be interesting to develop a strategy integrating the strengths of these two approaches to efficiently detect and to precisely delimitate ICEs and IMEs from Firmicutes.

References:

1. Bellanger, X., Payot, S., Leblond-Bourget, N., & Guédon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. FEMS microbiology reviews, 38(4), 720-760.

2. Cury, J., Touchon, M., & Rocha, E. P. (2017). Integrative and conjugative elements and their hosts: composition, distribution and organization. Nucleic acids research, 45(15), 8943-8956.

3. https://github.com/gem-pasteur/Macsyfinder_models/blob/master/Data/Conjugation/Tutorial_ICE.ipynb

4. Abby, S. S., Néron, B., Ménager, H., Touchon, M., & Rocha, E. P. (2014). MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PloS one, 9(10), e110726.

5. Abby, S. S., Cury, J., Guglielmini, J., Néron, B., Touchon, M., & Rocha, E. P. (2016). Identification of protein secretion systems in bacterial genomes. Scientific reports, 6, 23080.

6. Ambroset, C., Coluzzi, C., Guédon, G., Devignes, M. D., Loux, V., Lacroix, T., ... & Leblond-Bourget, N. (2016). New insights into the classification and integration specificity of Streptococcus integrative conjugative elements through extensive genome exploration. Frontiers in microbiology, 6, 1483.

Keywords: Integrative Conjugative Element, ICE, Integrative Mobilizable Element, IME, Mobilome, Bacterial Genomics, Comparative Genomics

Characterizing transcription factor combinatorics in cis-regulatory regions with supervised classification and sparse encoding

Quentin Ferré *† 1,2, Salvatore Spicuglia 1, Jacques Van Helden 1, Denis Puthier‡ 1, Cécile Capponi§ 2

1 Theories and Approaches of Genomic Complexity (TAGC) – Aix Marseille Université, Institut National de la Santé et de la Recherche Médicale : U1090, Centre National de la Recherche Scientifique – Théories et approches de la complexité génomiqueParc scientifique de Luminy163 avenue de Luminy13288 Marseille cedex 9, France

2 Laboratoire dÍnformatique et Systèmes (LIS) – Aix Marseille Université : UMR7020, Université de Toulon : UMR7020, Centre National de la Recherche Scientifique : UMR7020 – Aix Marseille Université – Campus de Saint Jérôme – Bat. Polytech, 52 Av. Escadrille Normandie Niemen, 13397 Marseille Cedex 20, France

Introduction

Transcription factors (TFs) are a class of regulatory proteins that bind to DNA on regions called CRE (cis-regulatory elements), so as to influence the transcription of a target gene. It is now understood that TFs work in combination, by competing and/or collaborating and forming complexes (Chaudhari et al., 2018). TF binding can be studied in silico through the prediction of Transcription Factor Binding Sites (TFBS) using Position-Weight Matrices (PWM, Mysickova et al, 2012). However only a fraction of predicted TFBS translate to actual binding sites. Another possibility is to use ChIP-Seq experiments (Chikina et al., 2012).

The combinatorics of TFs (their combined interactions) are often studied through statistics. Most works attempt to find co-occurring TFs pairs, i.e. pairs of TFs whose binding sites are often found in a closer proximity than would be expected by chance (Zhu et al., 2005). Other methods include unsupervised mining of association rules (Teng et al., 2014), finding TFs with correlated nucleosome occupancy (Lai et al., 2014), pointwise mutual information (Meckbach et al., 2015), and hypergeometric probability of occurrences (Terada et al., 2013). But as a whole, existing approaches actually seek TFs associations regardless of the type of CRE), and tend to study pairwise associations instead of n-wise combinations.

TF combinatorics are also of interest to CRE detection (Kleftogiannis et al, 2016), sometimes when combined with histone marks; for example, a software tool called TFcoop predicts a region's cis-regulatory activity using a suite of PWM matrices' scores (ie. nucleotide composition) for the region as variables in a Lasso regression (Vandel et al. , 2015). While these methods focus on CRE detection and annotation, they often consider each TF (and/or chromatin mark)

[*] Speaker

[†] Corresponding author: quentin.ferre@inserm.fr

[‡] Corresponding author: denis.puthier@univ-amu.fr

[§] Corresponding author: Cecile.Capponi@lis-lab.fr as an individual variable, rank them by importance, without considering the combinatorics of TFs.

Our objective is then twofold. First, we focus on to detecting TFs that are found associated to one each other. Second, we wish to uncover combinations of TFs that are characteristic of a class of CRE as opposed to other classes. We showcase our approach for different meanings of what a ”class” is : whether the different classes are different natures of CRE (enhancers vs promoters), or are of same nature but with different activities (active vs inactive enhancers). We propose a machine learning approach where an example is one cis-regulatory element, each characterized by a vector of features with each feature being the fixation level of one known TF as determined by ChIP-Seq.

Methods

We use three datasets focusing on three different kinds of biological problems in the K562 cell line, respectively : TFs combinations characterizing active enhancers, TFs combinations characterizing promoters that also exhibit enhancer activity, and a general application on all types of cis-regulatory regions using public data from ENCODE.

The first dataset was generated in our laboratory as part of the study of TF-based regulatory networks in developing primary thymocytes, using the p5424 cell line model. In this work, CRE were selected by computing the overlap between DHS (DNAse-I Hypersensitivity Sites) and ChIP-Seq peaks for 6 specific TFs. These regions were then assessed for enhancer activity using CapSTARR-Seq (Vanhille et al., 2015). Regions were then classified in three categories proportional to tagged activity ; unsupervised clustering using k-means was performed according to TFs fixation, proportional to mean ChIP-Seq signals around the region’s center (± 1kb for TFs, ± 5kb for histones).

The second dataset is based on a systematic CapSTARR-Seq analysis of E-promoters (Dao et al., 2017). E-promoters are promoters that also exhibit distal enhancer activity. For every human promoter, enhancer activity was assayed and a vector of TF fixation was quantified by the same method as above.

The third dataset is created using publicly available data (ENCODE Consortium, 2012), with ChromHMM prediction of genomic regions combined with ENCODE/HAIB ChIP-Seq TF peaks in the K562 cell line. We considered a bin for each region of 4kb around its center. For each region, we built a vector where each component corresponds to a score for a given TF ; that score is equal to the proportion of the bin covered by a peak for the given TF multiplied by the score of the peak. Scores are then L2-normalized.

To highlight class-specific profiles, we use decision trees as a clustering tool. A decision tree (Chen et al., 2007) is a model that aims at grouping samples in various nodes based on several input variables. Each leaf represents a ”cluster” which is as pure as possible (only composed of a single class whenever possible given the sample) given the values of the input variables represented by the path from the root to the leaf. The decision tree is used to perform a complex, combinated partitioning of the dataset. Unlike regular k-means clustering, this approach is supervised, allowing us to find class-specific profiles. Furthermore, different paths (with vastly different average profiles) can lead to nodes that are pure in the same class, highlighting diversity. Node splitting is performed by entropy and classes are rebalanced through oversampling. Since the decision paths only show variables that best discriminate between the classes, TFs correlated to a discriminative one will not be visible on the decision path, so we compute the average TF profiles across all the samples in each given node/cluster, allowing us to use a ”discriminative” decision tree as a clustering tool. For each node, class enrichment is computed using the hypergeometric law.

This first approach is compared to a sparse encoding of all the regions’ vectors computed via dictionary encoding : this approach rests upon the assumption that a matrix (here, our concatenated vectors) can be approximated by a sparse linear combination of special vectors called ”atoms” or ”words”, and seeks to find TFs combinations that are common across the entire dataset of studied CRE. Each line (or column) of the query matrix will be expressed as a combination of a single word in the dictionary, and a multiplicative coefficient (Li et al., 2012) Then, for each word in the dictionary, we analyse its usage and associated coefficients by class.

Results

On the first dataset, using our supervised classification method, we highlight complex interplay between different proportional fixations of Ets1 and Heb resulting in different enhancer activations. We also highlight the possibility of active enhancers lacking the H3K27ac histone mark, challenging the conventional view about its ubiquity (Creyghton et al., 2010). Dictionary analysis was used to study TFs combinations by class, meanwhile it analyses the k-means clusters previously computed. We show that there is a strong diversity of profiles per class, and that k-means clustering conceals this diversity; indeed k-means clusters enriched in Strong enhancers were found to have a similar, rather composite word usage profile.

Concerning E-promoters, given that they are usually active promoters, we compare them to a control set of promoters with equivalent activity : otherwise, we would have separated active and inactive promoters, not promoters and E-promoters. The decision tree structure is found to be onion-layered, with small, particularly class-enriched groups ”peeling off” from the bulk at each step. Previous analysis by et Dao al. (2017) showcased TFs enrichment for E-promoters, but only for each individual TF. In our work focusing on TFs combinations, we find that although many E-promoters have an EP300 and JUN-rich profile, a distinct subset is enriched in YY1 instead. There exists minor variations on these profiles that we dubbed ”accents”.

We are currently working on the ENCODE dataset. Unsupervised k-means clustering results in very impure clusters that do not exhibit different profiles, mostly grouping together regions with respectively high and low total TF fixation, although active promoters and insulators tend to regroup into a cluster of their own. It should be noted that there is a considerable number of regions for each class completely lacking in TF peaks; those regions are removed from the analysis. Further analysis is pending.

Conclusions

Our work allows us to highlight the diversity of TFs combinations profiles found within and between classes of cis-regulatory elements. It is a heuristics-based method, which can identify TFs tuples of arbitrary length. We discover both new and complex TFs combinations, but also reveal those to be characteristic of the CRE class they are found in. We are now looking to apply our method to the dataset compiled by (Muerder et al., 2018) which presents a whole-genome STARR-Seq, in order to experimentally evaluate enhancer activity across the human genome, as opposed to prediction by ChromHMM.

Our next endeavor will focus on the identification of Cis-Regulatory-Modules (CRM) using a deep learning approach. Previous work (DanQ, Quang et al., 2015) used deep learning with convolutional filters (CF) to classify genomic regions as enhancer or promoters, and found out that the CFs spontaneously learned correspond to many known TFBSs. We shall use a similar approach based on the distribution of ChIP-Seq TF peaks, considering for each position in the genome the presence of absence of a TF peak instead of its nucleotide (like DanQ). Then we shall analyze the filters learned by our model. A Long Short Term Memory layer should allow us to integrate positional dependencies.

References

Pedregosa et al., " Scikit-learn : Machine Learning in Python " JMLR 12, pp. 2825-2830, 2011.

Dao, Lan T. M., Ariel O. Galindo-Albarrán, Jaime A. Castro-Mondragon, Charlotte Andrieu-Soler, Alejandra Medina-Rivera, Charbel Souaid, Guillaume Charbonnier, et al. " Genome-Wide Characterization of Mammalian Promoters with Distal Enhancer Functions ". Nature Genetics 49, no 7 (juillet 2017): 1073-81. https://doi.org/10.1038/ng.3884.

Kleftogiannis, Dimitrios, Panos Kalnis, et Vladimir B. Bajic. " Progress and challenges in bioinformatics approaches for enhancer identification ". Briefings in Bioinformatics 17, no 6 (novembre 2016): 967-79. https://doi.org/10.1093/bib/bbv101.

Vanhille, Laurent, Aurélien Griffon, Muhammad Ahmad Maqbool, Joaquin Zacarias-Cabeza, Lan T. M. Dao, Nicolas Fernandez, Benoit Ballester, Jean Christophe Andrau, et Salvatore Spicuglia. " High-Throughput and Quantitative Assessment of Enhancer Activity in Mammals by CapStarr-Seq ". Nature Communications 6 (15 avril 2015): 6905. https://doi.org/10.1038/ncomms7905.

Vandel, Jimmy, Oceane Cassan, Sophie Lebre, Charles-Henri Lecellier, et Laurent Brehelin. " Modeling Transcription Factor Combinatorics in Promoters and Enhancers ". BioRxiv, 2 octobre 2017, 197418. https://doi.org/10.1101/197418.

ENCODE Project Consortium, " An integrated encyclopaedia of DNA elements in the human genome " Nature 2012 Sep 6;489(7414):57-74. https://doi.org/10.1038/nature11247

Vanhille, Laurent, Aurélien Griffon, Muhammad Ahmad Maqbool, Joaquin Zacarias-Cabeza, Lan T. M. Dao, Nicolas Fernandez, Benoit Ballester, Jean Christophe Andrau, et Salvatore Spicuglia. " High-Throughput and Quantitative Assessment of Enhancer Activity in Mammals by CapStarr-Seq ". Nature Communications 6 (15 avril 2015): 6905. https://doi.org/10.1038/ncomms7905.

Zhu, Zhou, Jay Shendure, et George M. Church. " Discovering functional transcription-factor combinations in the human cell cycle ". Genome Research 15, no 6 (juin 2005): 848-55. https://doi.org/10.1101/gr.3394405.

Kreiman, Gabriel. " Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes ". Nucleic Acids Research 32, no 9 (2004): 2889-2900. https://doi.org/10.1093/nar/gk

Terada, A., M. Okada-Hatakeyama, K. Tsuda, et J. Sese. " Statistical Significance of Combinatorial Regulations ". Proceedings of the National Academy of Sciences 110, no 32 (6 août 2013): 12996-1. https://doi.org/10.1073/pnas.1302233110.

Li, Yifeng, et Alioune Ngom. " Fast Sparse Representation Approaches for the Classification of High-Dimensional Biological Data ". In Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on: 4-7 October 2012, 2012. https://doi.org/10.1109/BIBM.2012.6392688. Chen, Xiaoyu, et Mathieu Blanchette. " Prediction of tissue-specific cis-regulatory modules using Bayesian networks and regression trees ". BMC Bioinformatics 8, no Suppl 10 (21 décembre 2007): S2. https://doi.org/10.1186/1471-2105-8-S10-S2.

Keywords: cis, regulatory regions, transcription factors, combinatorics, machine learning, decision trees, dictionary learning

CRISPR-Cas++, a webserver including tools for identification and analysis of CRISPR arrays and associated Cas proteins.

David Couvin *† [1], Aude Bernheim [3,2], Claire Toffano-Nioche [1], Marie Touchon [3,2], Juraj Michalik [4,5], Bertrand Néron [6], Eduardo Rocha [3,2], Gilles Vergnaud [1], Daniel Gautheret [1], Christine Pourcel [1]

[1] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette, France – CNRS : UMR9198 – France

[3] CNRS, UMR3525, 25-28 rue Dr. Roux, 75015, Paris, France – CNRS : UMR3525 – France

[2] Microbial Evolutionary Genomics, Institut Pasteur, 25-28 rue Dr. Roux, 75015, Paris, France – Institut Pasteur de Paris, CNRS : UMR3525 – France

[4] Université Lille 1, CRIStAL, équipe Bonsai, Cité Scientifique Bat M3, 59655 Villeneuve d'Ascq Cedex. – Université Lille I - Sciences et technologies – France

[5] AMIBio, INRIA Saclay, Bâtiment Alan Turing, 1 rue Honoré d'Estienne d'Orves, 91120, Palaiseau, France – L'Institut National de Recherche en Informatique et e n Automatique (INRIA) – France

[6] Bioinformatics and Biostatistics Hub – C3BI, USR 3756 IP CNRS – Paris, Institut Pasteur, 25-28 rue du Dr. Roux, 75015, France – CNRS : USR3756 – France

1. Introduction

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) elements consist in a succession of 24-47 base-pair long CRISPR repeats separated by similarly sized unique sequences called spacers. Together with a cluster of genes called *cas*, CRISPR-Cas systems constitute a defence mechanism for Bacteria and Archaea against mobile genetic elements, such as phages or plasmids. These systems can be used in multiple applications in genetic engineering and strain genotyping [1,2]. We present here the CRISPR-Cas++ webserver, which offers tools for the discovery and management of CRISPR and Cas loci in submitted nucleotide sequences. A major development is that of CRISPRCasFinder (submitted to NAR Web Server Issue 2018) a software allowing the identification of both CRISPR arrays and Cas proteins. This program is an updated, improved, and integrated version of CRISPRFinder [3] and CasFinder [4] with freely available third-party software dependencies. CRISPRCasFinder can either be used online or as a standalone tool compatible with Linux (including Windows Subsystem for Linux) and MacOS systems. We will also introduce the novel CRISPRCas database, which includes CRISPR arrays and Cas genes identified in all assembled viral, bacterial and archeal genomes.

2. CRISPRCasFinder

The CRISPRCasFinder program requires input provided in the form of a (multi-)Fasta file. The application has no predefined input size limit and is only limited by the available computer memory. CRISPR array and Cas protein analyses are returned as .xls, GFF3, JSON,

[*] Speaker

[†] Corresponding author: david.couvin@i2bc.paris-saclay.fr

TSV and Fasta formatted files. Other optional files and global statistics on CRISPR arrays can also be recovered when using the standalone program (which includes a variety of options). CRISPRCasFinder includes: (i) an evidence-level (1 to 4) rating system,

(ii) prediction of the potential orientation, and (iii) an updated Cas protein detection and typing procedure using CasFinder [4].

3. CRISPRCasDatabase and other tools

The CRISPRCasDatabase currently under construction includes data from predictions of CRISPRCasFinder for available prokaryotic genomic sequences. The database, built on the model of the CRISPRdb database (http://crispr.i2bc.paris-saclay.fr/) [5] will associate taxonomic information and propose lists of validated CRISPR arrays, repeats, spacers and Cas systems. It will be automatically updated when new sequences are available. CRISPR arrays that have been definitely validated by experts will be highlighted. Users will be able to manage their own data with MyCRISPRCasdb.

4. Conclusion and perspectives

CRISPRCasFinder offers an expert identification of CRISPR arrays and *cas* genes and compares favorably with other similar programs. The availability of CRISPRCasFinder both online and as a standalone program, provides more flexibility to potential users. Further ongoing developments of the related database will provide new functionalities to users. The web server will also include CRISPRtionary and CRISPRcompar, two tools allowing the classification spacers (and building of a dictionary) within various CRISPR arrays, and identification of loci by comparison of flanking sequences from selected strains [6]. The web server should help in investigating the complex evolution of the CRISPR-Cas systems and the relationship between the prokaryotic cells and mobile genetic elements. CRISPR-Cas++ is freely accessible at https://crisprcas.i2bc.paris-saclay.fr/.

References

1. Barrangou, R. and Doudna, J.A. (2016) Applications of CRISPR technologies in research and beyond. Nat Biotechnol, 34, 933-941. 10.1038/nbt.3659.

2. Hille, F. and Charpentier, E. (2016) CRISPR-Cas: biology, mechanisms and relevance. Philos Trans R Soc Lond B Biol Sci, 371. 10.1098/rstb.2015.0496.

3. Grissa I, Vergnaud G, Pourcel C. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Research 35:W52–W57. doi: 10.1093/nar/gkm360.

4. Abby SS, Néron B, Ménager H, Touchon M, Rocha EP. 2014. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PLoS One. 9(10):e110726. doi: 10.1371/journal.pone.0110726.

5. Grissa, I., Vergnaud, G. and Pourcel, C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics, 8, 172. 10.1186/1471-2105-8-172. doi: 10.1186/1471-2105-8-172

6. Grissa I, Vergnaud G, Pourcel C. 2008. CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. Nucleic Acids Research. 36:W145–W148. doi: 10.1093/nar/gkn228

Keywords: CRISPR, Cas systems, Cas proteins, prokaryote defence mechanism, standalone program, web interface

Understanding the gene regulatory network that directs human monocytes to dendritic cells differentiation

Karen Nuñez * [1], Paulina Pozos , Jesús Sotelo , Aurélien Naldi [2], Pablo Gonzalez , Christian Molina , Monica Padilla , Darely Gutierrez , Salvatore Spicuglia [3], M. Santana , Morgane Thomas , Denis Thieffry [4], Alejandra Medina-Rivera†

[1] International Laboratory for Human Genome Research-UNAM (LIIGH-UNAm) – Mexico
[2] Institut de biologie de l'école normale supérieure (IBENS) – École normale supérieure [ENS] - Paris, Inserm : U1024, CNRS : UMR8197 – 46, rue d'Ulm, 75005 Paris, France
[3] Theories and Approaches of Genomic Complexity (TAGC) – Aix Marseille Université, Institut National de la Santé et de la Recherche Médicale : U1090, Centre National de la Recherche Scientifique – Théories et approches de la complexité génomiqueParc scientifique de Luminy163 avenue de Luminy13288 Marseille cedex 9, France
[4] Institut de Biologie de l'Ecole Normale Supérieure (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris – France

Introduction
The aim of this project is to determine the regulatory mechanisms that govern the differentiation of monocytes to dendritic cells (DCs) and integrate the regulatory information in a logical model for DC differentiation. In peripheric tissues, differentiation from monocytes into DCs occurs mostly inflammatory conditions (Villadangos and Schnorrer, 2007). This differentiation can be reproduced *in vitro* (Segura et al, 2013). Albeit progresses in the field of DC differentiation, we still lack a comprehensive understanding of the underlying signaling pathways and regulatory circuits involved.

Methods

We analyzed bulk ChIP-seq and RNA-seq of DCs and monocytes from the BLUEPRINT consortium (Stunnenberg et al, 2016). Regarding bulk RNA-seq analysis, we used the methods described in Law et al (2016). Our analysis pipeline for ChIP-seq includes quality control with FastQC tool, along with ENCODE QC and IDR analysis (Bailey et al, 2013). Peak calling was performed using MACS2 (Hung and Weng, 2017). Chromatin states were then defined using ChromHMM (Ernst and Manolis, 2012). RSAT tool Peak-motifs was used to identify transcription factor binding sites (Thomas-Chollier et al, 2011).

Results

*Speaker

†Corresponding author: amedina@liigh.unam.mx

Using the software GINsim (Naldi et al, 2018), we are currently building a logical model of the regulatory network controlling dendritic cell differentiation from monocytes. This model includes the pathways previously described for CSF-2 and IL-4 signalling, taking into account the cross-talks between these pathways. CSF-2 and IL-4 activate the Jak/Stat pathway, leading to activation of the transcription factors STAT3, STAT5, STAT6, IRF4, NFKB2 and CEBAα, all reported to be involved in monocyte to dendritic cell differentiation.

The results of ChIP-seq and bulk RNA-seq analysis will enrich our model, taking into account novel transcription factors and regulatory interactions. Model analyses will serve as a screening platform to select components and interaction to assess experimentally (Collombet et al, 2017; Rodriguez et al, unpublished data).

Conclusion

DCs play an important role in the activation of the adaptive immune response, making them excellent candidates for cancer treatments and vaccines development. The construction of a predictive dynamical model of monocytes to dendritic cell differentiation should enable a better understanding of this process, to identify clinically relevant intervention points and ultimately design novel therapeutic strategies.

Keywords: Dendritic cells, modeling, differentiation, ChIP, seq, RNA, seq

Deciphering functional annotation of multiple gene sets with MOTVIS.

Aarón Ayllón-Benítez [*†] [1,2], Fleur Mougin[‡] [2], Manuel Quesada-Martínez [3], Jesualdo Tomas Fernández-Breis [4], Romain Bourqui [1], Patricia Thebault[§] [1]

[1] Université de Bordeaux – LaBRI – France
[2] Université de Bordeaux – Bordeaux Population Health Center – France
[3] University Miguel Hernandez (UMH) – Spain
[4] University of Murcia – Spain

Introduction
Currently, the advances in omics technologies have opened new opportunities in a large range of biological applications. Such advances may include single-cell, RNA-SEQ or microarray approaches that facilitate expression profiling according to a phenotype or a cell type of interest. As an illustration, these gene profiles are crucial to address the complexity of immune signatures [1]. As these approaches generate a large amount of information, they require bioinformatics pipelines to be understandable by biologists.

In practice, the detection of gene signatures is carried out by applying statistical approaches or clustering. Such methods aim at grouping genes according to their expression levels [2]. Then, deciphering the biological roles of these gene sets becomes a major research challenge to better understand and investigate the biological processes that are involved.

A relevant example is given by the human immunome where each cell has to play a specific role in the immune response. Then, an extensive cell type analysis can be carried out by gene sets that are specifically expressed in each cell type, making use of their gene profiles. For example, a group of genes associated with natural killer cells may be related to the innate immune response, antigen processing, presentation, and cytotoxicity. Thus, annotating gene sets is crucial to: (i) elucidate the biological role of these specific cells and (ii) highlight their specificity. Moreover, making use of these results as a whole can lead to pertinent applications for inferring the role of new type of cells. Furthermore the gene signature of each cell type has to be contextualized with the other types.

The annotation stage consists in associating a gene to a term described in a controlled vocabulary (inferred from experimental or automatic methods) describing functions, pathways, diseases, interactions, etc. This information is stored in various knowledge sources that are continuously evolving.

[*] Speaker
[†] Corresponding author: aaron.ayllon-benitez@u-bordeaux.fr
[‡] Corresponding author: fleur.mougin@u-bordeaux.fr
[§] Corresponding author: patricia.thebault@labri.fr

Managing the large number of annotation terms associated with a gene set level is usually very difficult. To address this issue, statistical methods, called enrichment methods, have been proposed [2,3]. These tools show an important pitfall related to redundancy in the results [4], resulting from the lack or under-exploitation of semantic relations between terms. In order to solve that, structure knowledge like the ontologies are proposed. The most widely used biological ontology is the Gene Ontology (GO) that provides almost 45 000 terms describing gene roles according to three sub-ontologies: biological processes, molecular functions and cellular components.

Few bioinformatics tools use multiple knowledge sources and aim at decreasing the redundancy and/or quantity of annotation terms by making use of semantic relations between terms [3,4]. However, to the best of our knowledge, no tool addresses these two features combined with a visualization system to analyze together related gene sets. In this context, visualization techniques provide real added-value for the expert when dealing with the additional level of complexity resulting from the multiple sets. So far, such aspects have been partially used to present enrichment results. For example, g:Profiler [5] uses a simple heatmap showing the presence or absence of a term for a given gene in the set. ClueGO [6] provides a node-link visualization between terms sharing the same genes. REVIGO [4] displays results according to three types of visualization: treemap, node-link and space diagram. However, the options avaible are very limited for dealing with multiple gene sets, . Moreover, these tools provide interaction options in the visualization to allow a deep exploration of results. In such context, we recently proposed a prototype of visualization tackling these issues [7], called MOTVIS (MOdular Terms Visualizations).

In this summary, we presents improvements of the MOTVIS pipeline and apply it, to the analysis of signatures of different types of cells. This type of analysis is becoming more interesting and requires new solutions to explore the functional signature of compared expression results since the emergence of single cell sequencing.

Methods

The workflow consists of three main steps to compute gene set annotations plus the dedicated visualization system to examine the results .

First, gene sets are annotated using an enrichment approach. g:Profiler has been chosen for this task because its uses several annotation databases. This permits to combine complementary knowledge for enriching functional information about gene sets (as gene annotations may have been done at different cell organization levels).

The second step involves a lexical analysis to infer relations between terms coming from different sources in order to eliminate redundant terms (same information about the functional roles). To do so, the OntoEnrich framework [8] has been integrated in for associating annotation terms with GO terms by following the strategy:(i) decomposing annotations into words, (ii) searching groups of consecutive words that correspond to a GO term or any of its synonyms, and (iii) removing words included in other ones.

Because of the large size of GO, the third step selects only the most relevant terms that synthesize the functional information of the input gene sets. Then, the most informative parent terms of each GO term found at the previous step are recursively processed until the root term is reached. The selection of the most informative parent term is computed using the information content score proposed in [9]. Once the subgraph of GO is created, the structure is explored to identify the GO terms associated with the gene sets. When a term is associated to the same gene sets as its ancestors, the ancestors are removed.

The last step is to explore these multi-set annotation results, for which a visualization tool has been designed (see Figure 1). The chosen visual structure combines an indented tree (to interactively move across the hierarchy of the ontology) and a circular treemap. The colored visualization of circular treemap represents the different hierarchy between terms and take into account the various scales of biological information that go from general to specific information. The proximity of some circles (representing annotation terms) require to use a visualization technique based on colours. We chose and adapted the three-colors algorithm [10] for automatically assigning gradients of colors to nodes according to their neighborhood distance while preserving a comprehensive cognitive understanding of their relative inclusion. The algorithm uses a color space that is recursively divided into intervals of colors associated with a node and its children. Then, increasing/reducing the luminance/sharpness improves the perception of depth in the tree. This visualization allows to explore the annotation results thanks to interactions as zoom and pan in the circular treemap, or click to expand the branch in the indented tree. Actions performed on the circular treemap impact on the indented tree (and vice-versa). In the circular treemap, the leaf node (white color) represents a gene set, in which a barplot summarizes all the annotations of this gene set (represented as colored circles).

Case study

To demonstrate the efficiency and reproducibility of the pipeline, the signature profiling of different types of cells has been analyzed using the data from The immunome compendium of immune cell subpopulations [1].The authors isolated 28 subpopulations of innate and adaptive immune cells, including normal mucosa and colon cancer cell lines. Each cell type presents different transcriptional profiles that can be considered as gene sets.

By applying the g:Profiler tool, we obtained 323 annotations for 24 gene sets using a hierarchical filter proposed in the tool. 98 annotation would have been obtained for 16 gene sets if only GO enrichment would have been done This demonstrates the great advantage of using several sources to characterize a larger number of gene sets. After using the lexical mapping, 264 out of the 323 annotations were kept (the 59 remaining annotations were discarded because they could not be mapped to GO). Five out of the 24 gene sets were ignored by our pipeline. Then, the hierarchy simplification stage (third stage) making use of the GO structure has decreased the number of annotations from 264 to 119. This 2.2-fold decrease demonstrates that the enrichment produces a significant quantity of redundant information.

Figure 1. Global view of the visualization tool. At this level, the global information that is displayed allows to define the three ontologies of GO (orange circle for biological process, purple for molecular function and blue for cellular component). The inclusive colored circles correspond to annotation terms that are included in the previous ones. At last, gene sets are represented as white circles.

To illustrate an application of MOTVIS (see a global view of MOTVIS in Figure 1), focusing on the cellular activation and migration, the indented tree can be interactively used to localize these annotation terms (Figure 2). They fall within "cellular process" and appear there as direct children of this general term (due to the simplification stage). Going into details within the "cell activation" circle, more specific annotations can be depicted. Moreover, if the "activation to lymphocyte" is the focus, zoom facilities are provided to identify the specific type of involved cells, in this case, T cells. At the leaf level (white circle related to T cells), the other annotations related to the type of focused cells can be observed. The pertinence of all the annotations is provided in the white circle thanks to the barplot (Figure 3).

Figure 2. Zoom in on "cell activation" and "cell migration" annotation terms. It shows the gene sets concerned by these annotations. The gradient of colors is correlated to the depth of terms within GO.

Figure 3. Zoom in to represent the leafs or white circles that are related to a type of cells. In this example, the information for the T cells is displayed. For this type of cell, all the annotation terms (corresponding to the gene profile) are represented within a barplot.

Conclusion

In this work, we present and apply the pipeline MOTVIS, dedicated to the annotation of multiple gene sets. Taking advantage of enrichment analysis and the use of several source knowledge, MOTVIS provides computation stages to: (i) perform an original lexical mapping that enables to make use of different knowledge sources, (ii) reduce the annotation redundancy and (iii) filter out the most relevant annotation to synthesize the functional information summarizing multiple gene sets. This new original pipeline has been applied to analyze, compare and visualize the results of a reference compendium of immune cells. According to the transcriptomic profiles of each cell type, MOTVIS offers an interactive way for both identifying the main roles where a type of cell may be involved, and deciphering common features between different cell types. According to the hierarchical relations between GO terms, biology experts can also choose the appropriate level of information (details on demand by interacting with the visualization system) to analyze the results.

BINDEA, G. et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity, 2013, p. 782-795.

HUANG, DW. et al. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 2008, p. 1-13.

THEBAULT, P. et al. Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks. Briefings in bioinformatics,

p. 795-805.

SUPEK, F. et al. REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one, 2011, vol. 6, no 7, p.

REIMAND, J. et al. g:Profiler-a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic acids research, 2007

BINDEA, G. et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics, 2009, p. 1091-1093.

AYLLÓN-BENITEZ, A. et al. Deciphering gene sets annotations with ontology based visualization. International conference in Information Visualization. 2017.

QUESADA-MARTÍNEZ, M. et al. Ontoenrich: A platform for the lexical analysis of ontologies. In: International Conference on Knowledge Engineering and Knowledge Management. Springer. p. 172-176.

ZHOU, Z. et al. A new model of information content for semantic similarity in WordNet. In: Future Generation Communication and Networking Symposia, 2008. p. 85-89.

TENNEKES, M. et al. Tree colors: color schemes for tree-structured data. IEEE transactions on visualization and computer graphics, 2014, p. 2072-2081.

Keywords: Visualization, Gene Ontology, functional annotation

Identification of regulatory variants involved in the development of sepsis

Florian Rosier * [1], Lydie Pradel [1], Pascal Rihet [2]

[1] Theories and Approaches of Genomic Complexity (TAGC) – Aix Marseille Université, Institut National de la Santé et de la Recherche Médicale : U1090, Centre National de la Recherche Scientifique – Théories et approches de la complexité génomiqueParc scientifique de Luminy163 avenue de Luminy13288 Marseille cedex 9, France

[2] Technologies avancées pour le génôme et la clinique (TAGC) – Inserm, Université de la Méditerranée - Aix-Marseille II – Parc scientifique de Luminy - 163 avenue de Luminy 13288 Marseille cedex 9, France

Since the advent of the Genome-Wide Association Studies (GWAS), a significant amount of Single-Nucleotide Polymorphism (SNP) has been identified as being associated with a phenotype or pathology. These SNPs constitute "tags" because hundreds of variants are inherited at the same time. Furthermore, many variants are in linkage disequilibrium (LD) with the SNPs associated with pathology. Firmly establishing the causality of a regulatory variant clearly requires the demonstration of molecular functionality and the identification of its (their) target gene(s).
Using a GWAS analysis, we identified 139 SNPs associated with early (before 7 days) or later (7 to 28 days) deaths phenotype in sepsis. Using bioinformatics analysis as impact of SNPs on transcription factor binding sites, eQTL analysis and region enhancer study, 2 regions of interest have been identified. These regions contain potentially one or more causal SNPs associated with death phenotype. The first region is located on chromosome 3 close to genes MAPKAPK3, CISH, HEMK1 or even DOCK3. These genes are particularly interesting because implicated in the immune response or cell differentiation. The second region is located on chromosome 9 close to genes RNF135 and ADAP2 known to be involved in the immune response against viruses.
We deleted each region using Cripr genome editing in the myeloid K562 cell line stimulated or not by the addition of lipopolysaccharide. The effect of deletion on genes' expression was then tested by qRT-PCR. In a second time, we will test each region SNPs one by one or in combination to precisely identify the causal variants involved in sepsis' development.

Keywords: SNP, GWAS, Crispr, Enhancer

[*] Speaker

Simulation of RNA sequencIng with Oxford Nanopore Technologies

Camille Marchet *† [1], Leandro Ishi * ‡ [2]

[1] Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) – Universite de Rennes 1, Institut National de Recherche en Informatique et en Automatique – Avenue du général LeclercCampus de Beaulieu 35042 RENNES CEDEX, France

[2] Laboratoire de Biométrie et Biologie Evolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 Bld du 11 Novembre 1918 69622 VILLEURBANNE CEDEX, France

Background

Oxford Nanopore Technologies (ONT) platforms produce reads order of magnitudes longer than previous generation of sequencers such as Illumina (several tens to thousands of base pairs). The unprecedented long range information these reads provide is currently used to better solve genomic assembly. However, these sequences suffer from high error rates (commonly above 10%) and complex error profiles: insertions and deletions are more common than in short reads and homopolymer regions are plagued by non stochastic errors due to the difficulty to estimate their length using ONT technology.

An increasing set of bioinformatic tools aim at making the best of these reads, in particular in assembly field but put also efforts at correction/polishing of the erroneous sequences. In any of these applications, long reads lead to algorithmic challenges due to their spurious nature, thus novel methods are propose to cope with their limitations. Several simulation tools, such as NanoSim to cite one of the most recent, were proposed to help benchmark novel algorithms on controled synthetic data, though as close as possible to real scenarios.

Until recently, transcriptomics applications with long reads were exclusively realized with Pacific Biosciences' protocol Iso-seq, recent advents of ONT moved forward from the genomic application to several transcriptomic protocols. Pioneer works start dealing with characterization of isoforms or gene expression quantification using nanopore reads. A few pipelines dedicated to transcriptomics studies appeared recently in the literature, such as alternative isoforms detection using ONT reads with Mandalorion pipeline. Indeed, ONT protocols are very promising in the transcriptomics context since they allow to discard PRC amplification step. Moreover a novel protocol allows to sequence directly from RNA molecules instead of relying cDNA templates. Thus ONT seem to avoid many previous biases encountered with Illumina, and they do not share Iso-seq's drawbacks that introduce bias in the transcripts representation that makes quantification step currently impossible. Finally, recent work is committed to detect RNA modifications such as methylations directly through ONT basecallers. However, ONT is not a technology as stable as Pacific Biosciences and proposed several successive chemistries in the last few months.

[*] Speaker

[†] Corresponding author: camille.marchet@irisa.fr

[‡] Corresponding author: leandro.ishi@univ-lyon1.fr

A current lack is the possibility to adequately simulate long reads from ONT RNA protocols, which would help with the developments of new tools to handle this kind of data. As pointed out in the past by RNA-seq simulators, the simulation of transcriptomics sequencing is a more complex task than in genomics because the gene expression and transcript variability have to be modeled. In this work we aim at filling this gap by proposing the first ONT RNA long reads simulator.

Contribution

Several features are required for an adequate simulation. First, gene and transcripts levels should be carefully computed in order to reproduce a biologically sound scenario. Since ONT is still fast evolving, the simulator should be able to adapt to the successive chemistries. Finally it should correctly render the reads common feature so that synthetic versions and real raw reads have close characteristics.

Our method is a pipeline that can be divided into four steps. For several of these steps, we relied on well established tools that we articulated with each other. It takes as input .BAM and .BAI from genome alignment as a training read set, .FASTA and .GTF of a reference genome and the desired final quantity of molecules.

The first module builds the error model for reads. As previously mentioned it is difficult to fit distribution for read errors, in particular since they change according to chemistries. Simulator such as Nanosim made the choice to learn error rates and profiles by training using real read data sets, and so do we. Alignments of reads from a real experiment the user wishes to mimic are passed as input, then the first module automatically deduces error rates and percentages of deletion, insertion and substitution, as well as homopolymer errors, using AlignQC. Pre-computed error profiles can also be input. The second module extracts transcripts that will be templates for the long reads from the GTF file of the desired reference using gffreads (http://ccb.jhu.edu/software/stringtie/gff.shtml). The third module is the expression levels definition. In order to simulate expression for transcripts of the input reference, we selected the Flux Simulator that enables detailed gene expression simulation. We use expression levels to decide which quantity of reads are generated per reads. The last module is a novel and efficient implementation in C++ that generates the final reads as well as their errorless versions for comparison matters. It adds the errors at positions in the sequences extracted from the GTF, deals with regular versus homopolymer errors and adds supplementary characteristics such as the staircase effect that affects length distribution, commonly encountered in this type of data.

This pipeline is a work in progress. Using mouse transcriptomic data, we will show the properties of our simulation and compare them to available mouse transcriptome ONT reads sequenced on the MinION platform at the Genoscope. Our method is written in Python and C++ and the sources and binaries are available on demand, and will be soon released on GitHub under open source license.

Conclusion

Transcriptomics studies using ONT protocols start to emerge as this technology harbours promising features that should help the identification and better comprehension of the modular nature of RNA variants, in particular in eukaryotes. Indeed, ONT long reads both give access to long range information about exon connectivity and benefit from protocols that reduce the biases in comparison to other sequencing technologies.

However only a few methodological contributions exist to date to handle this type of data. By providing the first tool for RNA ONT reads simulation we hope to help developers to assess their methods, and in particular enable future de novo softwares benchmarks.

Keywords: transcriptomics, oxford nanopore technologies, long reads, simulation, mRNA

Automatically identifying eu-hetero-chromatin boundaries through recombination rate estimates

Yasmine Mansour *† [1,2], Annie Chateau [1,3], Anna-Sophie Fiston-Lavier [2]

[1] Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier : UMR5506, Centre National de la Recherche Scientifique : UMR5506 – 161 rue Ada - 34095 Montpellier, France

[2] Institut des Sciences de lÉvolution de Montpellier (ISEM) – Université de Montpellier, Institut de recherche pour le développement [IRD] : UR226, Centre National de la Recherche Scientifique : UMR5554 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France

[3] Institut de Biologie Computationnelle (IBC) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut National de la Recherche Agronomique, Institut National de Recherche en Informatique et en Automatique, Université de Montpellier, Centre National de la Recherche Scientifique – 95 rue de la Galéra, 34095 Montpellier, France

Meiotic recombination is a vital biological process which plays an essential role for investigating genome-wide structural as well as functional dynamics. Various methods for estimating recombination rates exist in the literature. Population genetic based-methods [Stumpf and McVean, 2003] provide accurate fine-scale estimates. Nevertheless, these methods are very expensive, time-consuming, require a strong expertise and, most of all, are not applicable on all kinds of organisms. Moreover, the sperm-typing method [Jeffreys et al., 2000], which is also extremely accurate providing high-density recombination maps, is male-specific and share the same experimental requirements as population genetic methods. On the other hand, a purely statistical approach, the Marey Maps [Chakravarti, 1991], could avoid some of the above issues based on other available genomic data : the genetic and physical distances. The Marey maps for recombination estimates consist on correlating, for the same chromosome, the physical map with the genetic map containing respectively physical distances and genetic distances for a set of genetic markers. Despite the efficiency of this method and mostly the availability of physical and genetic maps, generating recombination maps rapidly and for any organism is still challenging. Hence, the increasing need of an automatic, portable and easy-to-use tool.

Here, we propose an automated bioinformatic solution based on the Marey maps method in order to provide local recombination rate estimates for various organisms. Furthermore, our approach allows to determine the eu-hetero-chromatin boundaries along chromosomes. This functionality is fundamental for identifying the location of the peri/centromeric and telomeric regions known to present a reduced recombination rate in most genomes. Most importantly for genomes which are provided as whole chromosomes instead of two arms per chromosome. We implemented our recombination tool by fitting a third-order polynomial to each chromosome based on genetic and physical maps. Compared to previous tools [Fiston-Lavier et al., 2010, Rezvoy et al., 2007], we have add a couple of new modules as to assess the quality of the data

*Speaker

†Corresponding author: yasmine.mansour@umontpellier.fr

(i.e. number and distribution of the markers along the genome) and to remove low-quality data according to the user's preference. Our approach automatically re-adjusts estimates in regions with a depletion of fitness between the polynomial and the data to detect the eu-hetero-chromatin boundaries for centromeric and telomeric regions in order to keep the estimates as authentic as possible to the biological process. Identifying these boundaries allows investigating recombination variations along the whole genome which will help comparing recombination patterns within and between species, especially insects in our case.

Our approach for the eu-hetero-chromatin boundaries detection has been primarily validated with cytological results that are experimentally generated on the *Drosophila melanogaster* genome [Comeron et al., 2012]. Moreover, since the pipeline we are proposing is non-genome-specific, our study is efficiently portable on other model as well as non-model genomes for which both genetic and physical maps are available. We have started interpreting the results on the mosquito specie *Culex pipiens*. We estimated the recombination rate along this genome and identified the heterochromatin boundaries on its three chromosomes. Also, after annotating its TEs, we have analyzed the correlation between TEs and recombination patterns. As in *D. melanogaster*, we observed non-homogenous distribution for active TE families such as LINEs and MITEs. In *Cx. pipiens*, while LINEs are enriched in pericentromeric regions, MITEs exhibit a higher density in euchromatin. In an attempt to explain such distribution bias, we investigated the dynamics for these two TE families through a comparative genomic approach carried out on other insect genomes.

We find our preliminary results quite promising since the TE distribution patterns across genomes generally show enrichment in specific regions such as constitutive heterochromatic exhibiting low recombination and low gene density. Therefor, we aim to take advantage of genome-wide recombination landscape to seek an explanation to the cause/effect association between recombination rate and TEs.

Keywords: Recombination rate, heterochromatin, transposable elements, comparative genomics, bioinformatics

evoDRUM: a new framework to model the evolution of metabolic networks under non-balanced growth conditions

Ghjuvan Grimaud *† [1]

[1] BIOMATHEMATICA – BIOMATHEMATICA – Quartier Balestrino 20000 Ajaccio, Corse-du-Sud, France

Mathematical modeling of evolution has been formalized in different frameworks, from population genetics1 to quantitative genetics2 and evolutionary game theory3. Each of them emphasizes a specific aspect of evolution, *e.g.*, the genetic mechanisms of inheritance, to the detriment of others, such as the ecology. More recently, a set of techniques called Adaptive Dynamics-theory4,5 has been developed to describe the evolution of phenotypic traits at the organismic level in a specific eco-evolutionary context, linking ecological and evolutionary dynamics. Adaptive Dynamics models the long-term consequences of small mutations in the phenotypic traits through their effects on fitness and competition of different mutants with the resident (with the initial trait value). Although Adaptive Dynamics (AD) is a powerful approach to model trait evolution, the mechanistic basis of trait innovations leading to new phenotypes is not included in this trait-based approach.

At the same time, the rapid development of systems biology lead to genome-scale metabolic network reconstructions of a large variety of microbial species, ranging from bacteria to unicellular eukaryotes6,7. These networks contain all the metabolic reactions of an organism, their associated stoichiometry and the genes encoding each enzyme. Several metabolic modeling frameworks have been proposed8, allowing a mechanistic derivation of phenotypic traits such as growth rates of species in a defined nutrient environment or the rate of production of a given metabolite. However, the kinetic modeling of each reaction is needed to calibrate the large set of associated biochemical reaction rates, which is experimentally difficult9. To overcome this limitation, most of the modeling methods such as Flux Balance Analysis (FBA)10 and dynamic Flux Balance Analysis (DFBA)11 rely on a simplifying assumption, the Quasi-Steady-State-Approximation (QSSA), that the internal metabolites do not accumulate inside the microorganism12. This is a good approximation for constant, balanced-growth conditions13 that, however, does not hold in temporally varying conditions (*i.e.* non-balanced growth). Recently, a metabolic modeling framework for non-balanced growth conditions, the Dynamic Reduction of Unbalanced Metabolism (DRUM), has been developed and validated on a microalga growing in a chemostat14. With DRUM, metabolic networks are divided into sub-networks assumed to be at quasi-steady state and connected by accumulating metabolites. The sub-networks are drastically reduced in size using Elementary Flux Modes analysis (EFM)15 while keeping the core information. For example, DRUM reduces the simplified stoichiometric matrix of the microalga *Tisochrysis lutea*, from 157 internal metabolites and 162 reactions to 16 metabolites and

[*] Speaker

[†] Corresponding author: gm.grimaud@gmail.com

8 reactions, respectively.

Recently, a new field of Evolutionary Systems Biology (ESB) has emerged16,17. While still in development, ESB aims to study phenotypes as a result of evolving intracellular interaction networks. More specifically, some efforts have been made to use FBA in an evolutionary perspective18,19 while considering a larger space of possible metabolic reactions gathered from the literature19. However, these approaches do not include ecological interactions, such as competition of mutants and residents. **Here, we developed evoDRUM, a new Evolutionary Systems Biology mathematical framework combining the novel metabolic modelling under dynamic conditions (DRUM) with the eco-evolutionary modelling of trait evolution (AD).** EvoDRUM extends and modifies the idea of gathering the evolutionarily possible reactions by defining a large - ideally universal - mutation space in which evolution can proceed. In line with the Adaptive Dynamics framework, evolution is driven by a step by step mutant/resident invasion dynamics, with a defined mutation rate. The novelty of the proposed approach is that it investigates the metabolically explicit trait changes and evolution as a result of selection through competitive interactions of different phenotypes. First, we fully developed this framework and applied it to several simple metabolic networks, with several resources and temporally fluctuating conditions. Then, we applied it to the genome-scale metabolic network of *Escherichia coli* and finally, to the evolution of simple microbial communities.

1) *Development of a new eco-evolutionary metabolic modeling framework, evoDRUM.* We developed evoDRUM, with three key new features: first, the method couples metabolic modeling to characterize growth of both the resident and the mutant and Adaptive Dynamics4,5 to describe a step by step resident/mutant invasion dynamics, where a mutant with novel trait(s) appears in a resident population and may replace it or not through competition, depending on their relative fitness. Second, the core metabolic model is DRUM14 allowing to drastically reduce the size of the system while keeping its core complexity. Third, we defined a large - possibly universal - mutation space in which one or several reactions corresponding to mutations are randomly chosen to obtain a mutant. The mutation space can be defined using metabolic reactions and metabolites from species closely related to the organism studied, or by using a larger set of taxa.

To fully develop evoDRUM, several assumptions must be made (*e.g.*, automation of the subnetworks cutting in DRUM). The numerical implementation is straightforward, because all the necessary computational tools already exist (the systems biology COBRA toolbox20 and the EFM algorithm in Matlab). The model is applied to realistic but simple tool metabolic networks and compared to evoFBA18 in the first steps of evolution; while evoFBA represents competition between different phenotypes using dFBA, it cannot take into account metabolic accumulation and evolutionary innovations. We investigated the effect of the number of mutations19, the mutation rate and neutral mutations21 on the evolutionary outcome. We also investigated how nutrient co-limitation22 and environmental fluctuations change the outcome, compared to constant conditions23.

2) *evoDRUM application to genome-scale metabolic networks.* evoDRUM was used to study the evolution of several organisms for which genome-scale metabolic networks are available. We defined a multi-level mutation space with different degrees of universality, from the genus level to the whole set of known metabolic reactions and metabolites. In contrast to the only comparable study19, the mutation space does not need to be a functional metabolic network on its own, *i.e.* once the metabolic reactions are collected, no further filtering is needed (evolution will be in charge of it). The first target organism was *Escherichia coli*24, for which extended 'omics data' are available. Evolution proceeded in constant and fluctuating conditions. We investigated the conditions leading to the development of fermentation vs respiration strategies25 and the

diauxic metabolic shifts26. Also, we investigated the effect of Horizontal Gene Transfer (HGT), corresponding to the addition of big blocks of reactions, on the evolutionary outcome27,28. We looked at the changes in modularity and connectivity in *E. coli* metabolic networks during evolution29. Finally, we used our method to retrace the evolutionary history of *E. coli* using phylogenetic relationships and comparative genomics analysis30,31 and compare it to existing results.

3) *Application of evoDRUM to the evolution of a simple microbial community.* The final step of the study was to apply evoDRUM to the evolution of a simple microbial community, such as a two-species cross-feeding system. Such systems can naturally arise within evoDRUM, through the evolutionary branching of strains. More complex microbial communities are also tractable within evoDRUM (*e.g.* n-species cross feeding interactions).

References.

1. Crow, J.F., Kimura, M. (1970). An Introduction to Population Genetics Theory; Harper & Row: New York, NY, USA.

2. Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. Evolution 1976, 314–334.

3. Maynard Smith, J. (1986). Evolution and the Theory of Games; Cambridge University Press: Cambridge, UK.

4. Dieckmann U. and Law R. (1996). The dynamical theory of coevolution: A derivation from stochastic ecological processes. Journal of Mathematical Biology, 34, 579-612, 1996.

5. Geritz S., Kisdi E., Meszéna G. and Metz J. (1998). Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. Evolutionary Ecology, 12, 35-57.

6. Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., & Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. Nature Reviews Microbiology, 7(2), 129-143.

7. Oberhardt, M. A., Palsson, B. Ø., & Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. Molecular systems biology, 5(1), 320.

8 Lewis, N. E., Nagarajan, H., & Palsson, B. Ø. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nature Reviews Microbiology, 10(4), 291-305.

9. Heijnen JJ, Verheijen PJT (2013). Parameter identification of in vivo kinetic models: Limitations and challenges. Biotechnoly Journal, 8, 768–775.

10. Orth J.D., Thiele I. and Palsson B. Ø. (2010). What is flux balance analysis? Nature biotechnology, 28 (3), 245-248.

11. Mahadevan, R., Edwards, J. S., & Doyle, F. J. (2002). Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophysical journal, 83(3), 1331-1340.

12. Price, N. D., Papin, J. A., Schilling, C. H., & Palsson, B. O. (2003). Genome-scale microbial in silico models: the constraints-based approach. Trends in biotechnology, 21(4), 162-169.

13. Song H-S, Ramkrishna D (2009). When is the Quasi-Steady-State Approximation Admissible in Metabolic Modeling? When Admissible, What Models are Desirable? Ind Eng Chem Res 48: 7976–7985.

14. Baroukh C., Munoz-Tamayo R., Steyer J.P. and Bernard O. (2014). DRUM: A new framework for metabolic modeling under non-balanced growth. Application to the carbon metabolism of unicellular microalgae. PloS one, 9 (8), e104499.

15. Schuster S, Dandekar T, Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends Biotechnol 17: 53–60.

16. Soyer, O. S. (2012). Evolutionary systems biology (Vol. 751). Springer Science & Business Media.

17. Soyer, O. S., & O'malley, M. A. (2013). Evolutionary systems biology: what it is and why it matters. Bioessays, 35(8), 696-705.

18. Großkopf, T., Consuegra, J., Gaffé, J., Willison, J. C., Lenski, R. E., Soyer, O. S., & Schneider, D. (2016). Metabolic modelling in a dynamic evolutionary framework predicts adaptive diversification of bacteria in a long-term evolution experiment. BMC Evolutionary Biology, 16(1), 163.

19. Szappanos B., Fritzemeier J., Cs´orgo B., Lazar V., Lu X., Fekete G., Balint B., Herczeg R., Nagy I., Notebaart R.A. et al. (2016). Adaptive evolution of complex innovations through stepwise metabolic niche expansion. Nature communications, 7.

20. Schellenberger, J., Que, R., Fleming, R. M., Thiele, I., Orth, J. D., Feist, A. M., ... & Kang, J. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0. Nature protocols, 6(9), 1290-1307.

21. Wagner A. (2008). Neutralism and selectionism: a network-based reconciliation. Nat Rev Genet 9:965-74. PubMed

22. Klausmeier, C. A., Litchman, E., & Levin, S. A. (2007). A model of flexible uptake of two essential resources. Journal of theoretical biology, 246(2), 278-289.

23. Soyer, O. S., & Pfeiffer, T. (2010). Evolution under fluctuating environments explains observed robustness in metabolic networks. PLoS Comput Biol, 6(8), e1000907.

24. McCloskey, D., Palsson, B. Ø., & Feist, A. M. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Molecular systems biology, 9(1), 661.

25. Wortel, M. T., Bosdriesz, E., Teusink, B., & Bruggeman, F. J. (2016). Evolutionary pressures on microbial metabolic strategies in the chemostat. Scientific reports, 6.

26. Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabási A-L-L, et al. (2007). Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. Proc Natl Acad Sci USA. 104:12663–8.

27. Pál, C., Papp, B., & Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic

networks by horizontal gene transfer. Nature genetics, 37(12), 1372-1375.

28. Iwasaki, W., & Takagi, T. (2009). Rapid pathway evolution facilitated by horizontal gene transfers across prokaryotic lineages. PLoS Genet, 5(3), e1000402.

29. Pfeiffer, T., Soyer, O. S., & Bonhoeffer, S. (2005). The evolution of connectivity in metabolic networks. PLoS Biol, 3(7), e228.

30. Navlakha, S., & Kingsford, C. (2011). Network archaeology: uncovering ancient networks from present-day interactions. PLoS Comput Biol, 7(4), e1001119.
31. Patro, R., Sefer, E., Malin, J., Marçais, G., Navlakha, S., & Kingsford, C. (2012). Parsimonious reconstruction of network evolution. Algorithms for Molecular Biology, 7(1), 25

Keywords: Evolutionary Systems Biology, Elementary Flux Modes, Evolution, Metabolic Networks

Charting the functional regulatory landscape of human RBPs through protein-RNA interaction predictions.

Andreas Zanzoni *† [1], Lionel Spinelli [2], Diogo Ribeiro [2], Gian Gaetano Tartaglia [3], C. Brun [4]

[1] Theories and Approaches of Genomic Complexity, U1090, Inserm- Aix-Marseille Université (TAGC) – Institut National de la Santé et de la Recherche Médicale - INSERM, Aix-Marseille Université - AMU – Parc Scientifique et Technologique de Luminy Case 928 13288 Marseille cedex 9, France
[2] Theories and Approaches of Genomic Complexity, U1090, Inserm- Aix-Marseille Université (TAGC) – Institut National de la Santé et de la Recherche Médicale - INSERM, Aix-Marseille Université - AMU – Parc Scientifique et Technologique de Luminy Case 928 13288 Marseille cedex 9, France
[3] Center for Genomic Regulation (CRG-UPF) – C/ Dr. Aiguader, 88 08003 Barcelona, Catalonia, Spain, Spain
[4] Theories and Approaches of Genomic Complexity, U1090, Inserm- Aix-Marseille Université (TAGC) – Institut National de la Santé et de la Recherche Médicale - INSERM, Aix-Marseille Université - AMU – Parc Scientifique et Technologique de Luminy Case 928 13288 Marseille Cedex 9, France

Transcripts coding for functionally-related proteins can be bound by common regulatory molecules, such as RNA-binding proteins (RBPs) and/or non-coding RNAs, thus forming the so-called RNA regulons. For instance, in yeast, protein-RNA interaction mapping studies demonstrated that many RBPs bind with specificity mRNAs coding for proteins involved in the same biological process (e.g., ribosome biogenesis, chromatin architecture, oxidative phosphorylation) or that are cytotopically related (e.g., cell wall, endoplasmic reticulum, mitochondria). In mammalian cells, several sets of related mRNAs may be part of RNA regulons as well (e.g., histone mRNAs, transcripts involved in inflammation and in DNA damage response).
A deeper understanding of the pervasiveness of the post-transcriptional regulation of related coding transcripts is subordinate to the availability of experimentally verified protein-mRNA interaction data. Despite the development of high-throughput methods to detect RNA molecules bound by RBPs, such as RNA immunoprecipitation and CLIP-based techniques, the protein-RNA interaction space is still largely unexplored. In this context, large-scale computational prediction of protein-RNA interactions can improve our understanding of post-transcriptional regulation.

To achieve this, we infer the functional landscape of the post-transcriptional regulation mediated by the human RBPs, by assessing the RNA regulon theory at different levels of organization of the cellular processes. For this, we developed and applied an original large-scale approach to predict human cellular processes post-transcriptionally regulated by RBPs, using a combination of protein-RNA interaction predictions, protein-protein interaction networks and statistical analyses.

*Speaker

†Corresponding author: andreas.zanzoni@univ-amu.fr

In order to identify the cellular processes potentially regulated through the binding of RBPs, we first computed the interaction propensities of 877 experimentally identified human RBPs with a representative set of 13,984 mRNA sequences, covering _˜63% of the human protein-coding genes, using the *cat*RAPID *omics* algorithm (Agostini et al. 2013). Doing so, we predicted the largest human mRNA–RBP interaction network to date consisting of more than 12 million interactions, of which 3.2 million show high interaction propensity scores (*cat*RAPID score ≥ 50).

We further investigated the predicted mRNA-RBP interaction network to characterize the functional landscape of the 877 RBPs. For each RBP, we assessed the over-representation of its interacting mRNAs among the transcripts encoding proteins involved in a same biological process or pathway. These were taken from four datasets, representing different levels of organization of the cellular functions (collectively hereafter named ”functional units”): *(i)* protein macromolecular complexes from the CORUM database; *(ii)* functional modules detected in a human protein interaction network using the OCG algorithm (Becker et al., 2012); *(iii)* pathways described in the KEGG database; and *(iv)* pathways from the Reactome knowledgebase. For each functional unit, we have computed the ratio of interacting vs. non-interacting transcripts for each RBP and assessed its significance to be higher or lower than expected by chance by performing a two-sided Fisher's Exact test. In this way, we built a predicted regulatory landscape for 713 RBPs comprising 3185 significant enrichments involving 250 functional units as well as 2314 significant depletions involving 77 functional units (300 functional units in total).

The first important result of our analysis is the presence of functional unit enrichments, which represent putative post-transcriptional regulatory events, but also of depletions among RBP predicted targets. Indeed, the fact that _˜42% of the significant results are represented by depletions is intriguing, as it may suggest that some functional units may avoid RBP binding. Further scrutiny let emerge an interesting pattern of enrichment/depletions that allows grouping both RBPs and functional units in three broad categories: 1) a relatively small number of RBPs showing exclusively enrichments (75 RBPs, _˜10% of the RBP with significant results), 2) 211 RBPs displaying only significant depletions (_˜30%), and 3) 427 RBPs having both significant enrichments and depletions (_˜60%). Similarly, we observed that a majority of the functional units (223 functional units, 74% of the units with significant results) are exclusively enriched among the predicted targets of at least one RBP, a smaller proportion (50 functional units, _˜17%) show only significant depletions, and few functional units, namely 27 (9%), are both enriched and depleted in RBP predicted targets.

We found that RBPs classified as classic metabolic enzymes or as proteins lacking a recognized RNA-binding domain, have a significantly higher number of functional unit enrichments that canonical RBPs, thus indicating a more promiscuous regulatory potential for these unorthodox RNA-binding proteins. Moreover, we noticed that the predicted targets of 125 RBPs are enriched in units belonging to their own functional milieu (i.e., units that share a significant number of components with the cellular processes in which the RBPs are known to be involved), meaning these RBPs can bind functional neighboring transcripts and ensure a coordinated post-transcriptional regulation of their related cellular processes.
From a functional unit perspective, we labelled as ”frequently enriched” several cellular processes and components that are known to be regulated at the post-transcriptional level (e.g., autophagy, mitochondrial ribosome, energy-related pathways and histones). We also found several infection-related pathways to be enriched among the predicted interactors of dozens of RBPs, which we propose as potential RNA regulons. Finally, we observed a more tissue-specific expression of those functional units that are exclusively depleted. This result suggests that RBP-binding avoidance may be required for proper expression of tissue-specific functions.

Keywords: RNA, binding proteins, mRNA, protein, RNA interactions, protein, protein interaction network

IBENS Genomics core facility

Laurent Jourdren * [1], Corinne Blugeon [1], Fanny Coulpier [1], Berengere Laffay *

[1,2], Sophie Lemoine *

[1], Ammara Mohammad [1], Stéphane Le Crom [1,3]

[1] Institut de biologie de l'école normale supérieure (IBENS) – École normale supérieure [ENS] - Paris, Inserm : U1024, CNRS : UMR8197 – 46, Rue d'Ulm. 75005 Paris, France
[2] Master Bioinformatique, Normandie Université, UNIROUEN (UNIROUEN) – Université de Rouen Normandie – France
[3] Institut de Biologie Paris Seine – Sorbonne Université UPMC Paris VI – France

The **genomics core facility of the Institut de Biologie de l'École normale supérieure (IBENS)** [1,2] was created **in 1999**. We have been focused on **eukaryotes and specifically on functional genomics** analyses since the beginning. We handle **classical model organisms** and also more **exotic organisms** (jellyfish, birds, butterflies...). The **facility has always been a well-balanced structure between wet-lab and bioinformatics**: half of the team is involved on the wet-lab part; the remaining half being involved on the data analysis part. Our goal is to **help laboratories** during their **high-throughput projects from the experimental design to data analysis for publication**.
In 2008, we joined the **France Génomique** consortium, which has been financed by the governmental funding program "Investissement d'Avenir" since 2010. We have been following the **ISO 9001 quality international standard** since March 2013 and the **NF X 50-900** certification defined by IBiSA since April 2015.

Our activity **started in 1999 with DNA microarrays**. Facing the rise of high-throughput sequencing in 2010, we decided to **invest in a high-throughput sequencing machine**, an **Illumina HiSeq 1500** and we gradually moved from DNA microarrays to RNA sequencing, ending our support for biochips in 2013. The **HiSeq 1500** was then replaced by a **NextSeq 500** at the beginning of 2015. Since 2010, 3,020 samples have been sequenced on the facility.

All the staff working on the facility gets a balanced schedule between the core production service and research and development projects to propose **up to date and reliable experimental solutions** to our collaborators. To cope with the experimental constraints of our collaborators among the research teams (a lot of neuroscience and developmental biology teams), we invest a lot of our time in **testing library protocols** (very low quantities, ribosome depletions...). We are also **deeply involved in software development** to manage our project analyses (40% of projects are analysed on the facility). The tools we develop are distributed on an **open source** basis on GitHub [3] and we now provide most of them as **Docker images** [4] to **ease the**

*Speaker

distribution of our work. Our concern is to develop **workflows to achieve reproducible and transparent data analysis** of our high throughput experiments. We were among the first in France to **provide cloud computing data and big data analysis**.

Six of our software have been published in peer reviewed publications. **We currently develop and maintain three software**. The first one, **Aozan** [5] is an automated **post sequencing data processing pipeline** that automatically handle data transfer, demultiplexing conversion and quality control once an Illunima sequencer run is finished. The second one, **Eoulsan** [6] is a **versatile workflow manager that can reproducibly analyse huge amounts of sequencing data**. The last one, **ToulligQC** [7] is a software dedicated to the **QC analyses of Oxford Nanopore runs**.

We work on a collaboration partnership mode with the research teams. We are proud to participle in co-authorship of papers, **23 peer reviewed publications** during the last 4 years (2014-2017). We are also highly **involved in training,** we regularly participate to lectures and practical sessions for **students and research staff**.

Since 2016, we have been working on two main development projects.

The first one is devoted to **long read sequencing in RNA-seq**. We work with **Oxford Nanopore Technologies** (ONT) MinION system in order to sequence **full length transcript for isoform abundance estimation**. Thanks to our deep-rooted experience in RNA-seq and good results in ONT sequencing, we were **involved in ONT RNA dedicated protocol testing**. **We believe in a strong involvement in the ONT community**, we therefore participated to an online webinar in September, we were invited as a speaker to the **Nanopore Community Meeting** in New York last November and we hosted a **Nanopore Day in Paris** in March 2018. To go further and improve **knowledge sharing inside the French Nanopore community**, we are launching a 2-3 times a year **wet-lab and bioinformatical workshop**.

Our second development project is dedicated to **single cell RNA-seq**. We recently purchased a **Chromium** system from **10X Genomics** based on the **Drop-seq protocol**. We are currently working on the **validation of the protocols and new steps in Eoulsan**, our analysis pipeline. This validation work is performed in collaboration with Piotr Topilko from the Development of the nervous system team and Denis Thieffry from the Computational biology of the systems team. **Once fully validated, the single cell technology will be available to our collaborators**.

All these on-going projects allow us to be at the state of the art in functional genomics applications so that we can provide the Paris area scientific community all the tools needed to succeed in their high throughput experiments.

http://genomique.biologie.ens.fr

Twitter @Genomique_ENS

https://github.com/GenomicParisCentre/

https://hub.docker.com/r/genomicpariscentre/

http://outils.genomique.biologie.ens.fr/aozan/

http://outils.genomique.biologie.ens.fr/eoulsan/

https://github.com/GenomicParisCentre/toulligQC

Keywords: Genomics core facility, RNAseq, Long reads, Single cell RNAseq, Software development

Enrichment analysis with EpiAnnotator.

Yoann Pageaud *† [1], Christoph Plass [2], Yassen Assenov‡ [2]

[1] Division of Epigenomics and Cancer Risk Factors, German Center for Cancer Research (DKFZ) – Heildelberg, Germany
[2] Division of Epigenomics and Cancer Risk Factors, German Center for Cancer Research (DKFZ) – Heidelberg, Germany

MOTIVATION:
Deciphering relevant biological insights from genomic and epigenomic data can be a challenging task. One commonly used approach is to perform enrichment analysis. However, finding, downloading and using the publicly available functional annotations requires time, programming skills and IT infrastructure. We designed the online tool EpiAnnotator for performing enrichment analyses based on epigenomic and genomic data in a fast and user-friendly way.

RESULTS:

EpiAnnotator is an R Package accompanied by a web interface. It contains regularly updated annotations from 4 public databases: Blueprint, RoadMap, GENCODE and the UCSC Genome Browser. Annotations are hosted locally or in a server environment and automatically updated by scripts of our own design. Thousands of tracks are available, reflecting data on a variety of tissues, cell types and cell lines from the human and mouse genomes. Users need to upload sets of selected and background regions. Results are displayed in customizable and easily interpretable figures.

AVAILABILITY:

The R package and Shiny app are open source and available under the GPL v3 license. EpiAnnotator's web interface is accessible at http://computational-epigenomics.com/en/epiannotator.

CONTACT:
epiannotator@computational-epigenomics.com.

Keywords: Enrichment Analysis, Epigenomics, Genomics, Methylation, Cancer, R Package, Web Service, Blueprint, RoadMap, Shinny App

*Speaker

†Corresponding author: yoann.pageaud@gmail.com

‡Corresponding author: y.assenov@dkfz-heidelberg.de

FoodMicrobiome Transfert : Web based Metagenomic Analysis

Quentin Cavaillé [*] [1], Thibaut Guirimand[†] [2], Sandra Derozier[‡] [3], Anne-Laure Abraham[§] [2], Bedis Dridi[¶] , Valentin Loux[‖] [4], Pierre Renault[**] [5]

[1] MICrobiologie de l'ALImentation au Service de la Santé humaine (MICALIS) – Institut National de la Recherche Agronomique : UMR1319, AgroParisTech – F-78350 JOUY-EN-JOSAS, France
[2] MICrobiologie de l'ALImentation au Service de la Santé humaine (MICALIS) – AgroParisTech, Institut national de la recherche agronomique (INRA) : UMR1319 – F-78350 JOUY-EN-JOSAS, France
[3] INRA, UR1404 Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE) – Institut national de la recherche agronomique (INRA) – Bâtiment 210-233 Domaine de Vilvert 78350 Jouy en Josas Cedex, France
[4] Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE) – Institut national de la recherche agronomique (INRA) : UR1404 – Bâtiment 210-233 Domaine de Vilvert 78350 Jouy en Josas Cedex, France
[5] INRA-CRJ, Jouy-en-Josas – Michalis Institute – Michalis Institute, Département Microbiologie et Chaîne Alimentaire, INRA-CRJ, Jfouy-en-Josas, France, France

A large number of micro-organisms are involved in the composition of cheeses1 : bacteria, yeasts, filamentous fungi, phages. These micro-organisms come from starters used by manufacturers, but also from the environment (milk, maturing cellars, salt). In order to achieve a better understanding of these flora, metagenomic approaches can be used2 . The FoodMicrobiomes Trans-

fer project brings together industry and academia for the purpose of providing a tool to ease the analysis of metagenomic sequencing data, especially cheese flora samples, via a web interface. Cheese flora has long been studied, which ensure the availability of several hundred of reference genomes. The partners concerned by the project wish to identify organisms present in the ecosystem with a precise taxonomic assignation, down to the strain level (different strains may have different properties) and also identify micro-organisms present in low abundance. To an-swer

these questions, we chose a shotgun metagenomic sequencing approach, coupled with a tool that aligns the metagenomic reads on the reference genomes. This tool can be used via a web inter-face,

coupled with a database of around 4000 genomes from dairy exosystems and metagenomes. This tool also identifies, for each reference genome, which genes are present into the ecosystem.
The FoodMicrobiome web interface (http ://migale.jouy.inra.fr/foodMicrobiome/) allows users

[*] Speaker

[†] Corresponding author: thibaut.guirimand@jouy.inra.fr

[‡] Corresponding author: sandra.derozier@jouy.inra.fr

[§] Corresponding author: anne-laure.abraham@jouy.inra.fr

[¶] Corresponding author: bedis.dridi@inra.fr

[‖] Corresponding author: valentin.loux@jouy.inra.fr

[**] Corresponding author: pierre.renault2@inra.fr

to perform analyses on their own metagenomes on reference sequences (public or not). In order to guide the analyzes, predefined lists of genomes are available. Users can also create their own lists of interest for their own analyzes, or share them with other users.

The FoodMicrobiome application is based on a PostgreSQL relational database. This database allows the management of genomes, genome lists, metagenomes and analytical results. Computations are performed transparently for the user on the Migale platform's calculation cluster via the Bioblend API and the Galaxy portal. The web interface is developed via the Python Django framework as well as web technologies such as HTML and JavaScript.

The FoodMicrobiome application was designed and developed to study cheese ecosystems. It can be adapted to other ecosystems for which we have sequenced genomes, for example food ecosystems. It will be available on the migale plateforme.

1. Marie-Christine Montel, Solange Buchin, Adrien Mallet et al. "Traditional cheeses : Rich and diverse microbiota with associated benefits". In :International Journal of Food Microbiology 177 (2014) pp.136–154

2. Bhagya. R. Yeluri Jonnala, Paul L. H. McSweeney et al. "Sequencing of the Cheese Microbiome and Its Relevance to Industry". In :frontiers in Microbiology, doi : 10.3389/fmicb.2018.01020

Keywords: Metagenomics, Django, Galaxy, Python

ARTwork, a bioinformatics solution dedicated to bacterial WGS data management and analysis in the context of foodborne pathogens surveillance and outbreak investigations

Kevin Durimel *† [1], Arnaud Felten [1], Michel-Yves Mistou [1]

[1] French Agency for Food, Environmental and Occupational Health Safety [Maisons-Alfort] (ANSES) – Anses – 27-31 Av. General Leclerc 94701 Maisons-Alfort, France

With ever-increasing amounts of raw data produced by whole genome sequencing (WGS) projects, some typical bioinformatics processes (e.g quality control, genome assembly and annotation), are frequently and repeatedly performed. Final users in public health laboratories (scientists, engineers, technicians) receive a growing flow of bioinformatics data that must be processed in a reproducible way and must be easily retrievable. The traceability and reproducibility of genomic data analysis is particularly important in the context of official activities like outbreak investigations or epidemiological surveillance.

To guarantee the traceability of bioinformatics analyses, it is mandatory to keep a track of pre-sequencing and post-sequencing processes, like data and tools used, as well as computational steps and their parameter options. To ensure quick data recovery, a regularly saved database is a common solution. A multitude of software, tools, and cloud solutions already provide solutions to meet these requirements [1] [2] [3]. Nevertheless, they cannot be executed routinely, are not fully open-source and are not suitable for use in the context of foodborne pathogens surveillance. Therefore, we have designed an open source bioinformatics solution aiming to standardize WGS data analysis for the most common foodborne pathogens tracked in the laboratory (i.e *Bacillus, Clostridium, Listeria, Salmonella, Staphylococcus*).

This solution named *ARTwork* (Assembly of Reads and Typing WORKflow), is a workflow implementing 15 bioinformatics tools performing reads quality control, reads trimming, variant calling, de novo assembly, reference-based scaffolding, assembly quality control, genome typing and annotation, and tracking all metadata generated during the data processing lifecycle. A document-oriented database, MongoDB, is also used to store information about each WGS data received. Finally, a user-friendly web application allows final users to query the database and easily retrieve data which can be used as a new starting point for downstream analyses.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3414708/

https://www.ncbi.nlm.nih.gov/pubmed/20069535

[*] Speaker

[†] Corresponding author: kevin.durimel@anses.fr

http://www.applied-maths.com/bionumerics

Keywords: wgs, foodborne pathogens, outbreak investigations, data management, assembly, typing, variant, calling, database, interface

Galaxy-based Interactive ANalysis of Transcriptomic data (GIANT): A modular Galaxy pipeline as an alternative to licenced softwares for analyzing transcriptomic data

Jimmy Vandel [1], Céline Gheeraert [1], Jérôme Eeckhoute [1], Bart Staels [1], Philippe Lefebvre [1], Julie Dubois-Chevalier *† [1]

[1] Laboratoire Récepteurs Nucléaires, Maladies Cardiovasculaires et Diabète (U1011) – Université Lille Nord (France), Institut National de la Santé et de la Recherche Médicale - INSERM, CHU Lille ,, Institut Pasteur de Lille, U1011-EGID – Institut Pasteur de Lille 1, rue du Pr Calmette BP 245 59019 Lille Cédex, France

Context and rationale
Microarrays have been extensively used to analyze transcriptomes in functional genomic studies. While RNA-seq is becoming a new standard, microarrays are still used in specific instances and transcriptomic data obtained with this technology populate public databases such as Gene Expression Omnibus (GEO) [1].

While biologists would benefit from being able to interrogate these data, only proprietary tools allow for non-programming analysis of microarrays in a comprehensive manner on the same platform/tools-suite. While some R/Bioconductor packages allow to handle microarray data, in our knowledge, a user-friendly solution allowing to perform QC plots, normalization and complex differential analyses and interpretation plots (volcano and heatmap) at once is lacking. Analyses of RNA-seq is also still limited by the availability of user-friendly tools regarding data mining such as clustering.

Overview of GIANT

We have developed GIANT, a modular Galaxy tool suite interfacing R packages [2] (as Limma [3] and Oligo [4]) to allow non-programmers to perform transcriptomic analyses in a user-friendly environment. On the one hand, this suite allows users to consecutively perform all required steps required for microarray-based analyses ranging from data QC to differential analyses and complex visualization. On the other hand, specific modules such as QC-plot, limma-differential-analysis or heatmap-on-differential modules can be used with any type of pre-processed data. Each module provides the user with opportunities to adjust parameters for 'tailor-made' analyses. Furthermore, we use the power of Plotly [5] and Heatmaply [6] packages to provide interactive graphics and interactive requestable result tables in each module, a very innovative feature for Galaxy tools.

*Speaker

†Corresponding author: julie.chevalier@inserm.fr

Details of the modules of GIANT

Our tools-suite wraps functionalities for: 1) QC plot generation (Array images, boxplots, signal density plots, MA plots, PCA), data normalization (with limma [3] or APT tool from Affymetrix, with probe-set and gene-level analysis), 2) complex differential analyses with no restriction on number of conditions and factors and taking into account blocking effects (as batch effects or paired-analysis), 3) data visualization with raw p-values histograms, volcanos and circular plots and 4) heatmap and hierarchical clustering generation based on normalized expressions and on differential analysis results.

Interactive graphics integrated in our modules allow users, for example : to explore data within 3D PCA plots, to target particular points in the volcano plot to easily identify genes of interest or to zoom in heatmaps to explore the content of clusters.

Conclusion

By developing GIANT, a modular Galaxy tool suite, we allow non-programmers to benefit from the power of R/Bioconductor packages to perform a complete and complex analysis of transcriptomic data in a free, highly interactive, and user-friendly environment. This modularity allows this suite to be used not only for analysis of microarray data but also for interpreting RNA-seq data.

Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 ; 30(1):207-10

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015 ; 43(7),e47

Carvalho B. S., and Irizarry, R. A. A Framework for Oligonucleotide Microarray Preprocessing. Bioinformatics. 2010

Carson Sievert, Chris Parmer, Toby Hocking, Scott Chamberlain, Karthik Ram, Marianne Corvellec and Pedro Despouy. Plotly: Create Interactive Web Graphics via 'plotly.js'. 2017. R package version 4.7.1. https://CRAN.R-project.org/package=plotly

Tal Galili, Alan O'Callaghan, Jonathan Sidi, Carson Sievert. heatmaply: an R package for creating interactive cluster heatmaps for online publishing. Bioinformatics. btx657, https://doi.org/10.1093/bioinfor

Keywords: Transcriptomic, Galaxy, Interactive, tool suite, microarray, RNAseq

BIG: BioInformatics and Genomics platform at Institut Sophia Agrobiotech

Martine Da Rocha *† [1], Etienne G.j Danchin *

[1], Corinne Rancurel * ‡ [1]

[1] Institut Sophia Agrobiotech [Sophia Antipolis] (ISA) – Institut National de la Recherche Agronomique : UMR1355, Université Nice Sophia Antipolis : UMR7254, Centre National de la Recherche Scientifique : UMR7254 – INRA Centre de recherche Provence-Alpes-Côte dÁzur 400, route des Chappes BP 167 06903 Sophia Antipolis Cedex, France

Institut Sophia Agrobiotech is a joint research unit of INRA, CNRS and the University of Nice - Sophia Antipolis, mainly interested in biotic and abiotic factors influencing plant health. The laboratory uses molecular biology, biochemistry, population genetics, comparative genomics, evolutionary biology and modeling approaches to study the interactions with plants influencing their health. Currently, the institute is composed of 11 research teams and one core facility team named SPIBOC (for Plant Health, Biotic Interactions: Shared Common Scientific tools) that provides access to equipments, expertise and tools to support the research of the different teams. The SPIBOC team is directed by Karine Hugot (INRA research engineer) and made of 3 different platform: (i) Imagery and Microscopy (MIC), (ii) Biochemistry and Mass Spectrometry (BA), (iii) Bioinformatics and Genomics (BIG).
The BIG platform was the most recently created platform as a solution to cover the ever-growing demand of the different research teams in terms of services, expertise and skills in bioinformatics and genomics. Indeed, with the democratization of sequencing technology, the different teams of the Institute have been producing a growing amount of sequence data and this trend is continuing.

The core of the BIG platform is composed of two bioinformatics engineers: Martine Da Rocha (from INRA) and Corinne Rancurel (from CNRS). The core is complemented by a scientific advisor: Etienne Danchin (INRA senior scientist).

The BIG platform is open for collaboration and can be contacted at the following e-mail address: spiboc.big@inra.fr

A web page summarizes the activities and organization of the BIG platform:

http://www6.paca.inra.fr/institut-sophia-agrobiotech/Infrastructure-PlantBIOs/Plateforme-SPIBOC/Plateau
de-bioinformatique Or: http://tinyurl.com/y9qkho4v

*Speaker

†Corresponding author: martine.da-rocha@inra.fr

‡Corresponding author: corinne.rancurel@inra.fr

The BIG platform is integrated in the network of bioinformaticians from the Plant Health & Environment of INRA. This network called BBRIC (http://cati-bbric.toulouse.inra.fr/) encompasses 41 engineers from 12 laboratory across France.

The expertise of the BIG platform mainly lies in the field of biotic and abiotic interactions influencing plant health but the different skills and services proposed by the platform can be applied to other research topics.

So far, the main routine services proposed by the platform as semi-automated pipelines are :

- Differential gene expression from RNA-seq data
- Gene copy number variations detection from CGH array data.
- Detection of horizontal gene transfers

Besides these main routine services, the BIG platform regularly performs customized projects involving:

- De novo and genome-guided transcriptome assembly.
- Genome assembly (including from 3rd generation sequencing data)
- Genome annotation (including gene prediction and annotation of transposable elements)
- DNA methylome analysis using short bisulfite sequencing data
- SNP-calling from alignment of genomic reads to reference genomes.
- Identification of small RNAs from from specialized RNA-seq libraries.
- Functional annotation, including prediction of protein domains, signal peptides and assignment of Gene Ontology terms.
- Phylogenetic analyses
- OTU identification from metabarcoding data

Here are some examples of software or web portals developed by the BIG platform:

> Alienness (alienness.sophia.inra.fr), Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life, a web tool for the high-throughput detection of horizontal gene transfers.
>
> SATqPCR, (satqpcr.sophia.inra.fr), Statistical Analysis Tool for quantitative Real-time PCR Data, a web tool for the statistical analysis of quantitative PCR data. Meloidogyne genome resources , (meloidogyne.inra.fr), The web portal for the genomes of root-knot nematodes, the most devastating plant-parasitic worms.

The BIG platform is currently equipped with the following material:

- 1 PowerEdge R520 server, 32Go RAM, 3,5 To HDD, 12 CPUs

hosting 6 virtual machines

- 1 PowerEdge T630 server, 32Go RAM, 13 To HDD, 12 CPUs
- 1 PowerEdge R930 server, 765 Go RAM, 14 To HDD, 72 CPUs
- 1 PowerEdge R740 server, 32Go RAM, 70 To HDD, 16 CPUs

As other resources, we have access to the following bioinformatics plateforms : genouest bioinformatics and genotoul bioinformatics.

In addition to its activity in development of software, treating projects and collaborating with research teams, the BIG platform also dedicates a large part of its time to the transmission and sharing of knowledge by:

- Supervising and co-supervising french and international students (mainly Masters and PhD)
- Organizing practical informatics and bioinformatics courses for scientists
- Communicating in science exhibition fairs

In terms of scientific production, on the period 2012-2017, the BIG platform has been involved as co-author in 17 scientific publications in peer-reviewed journals, including in Nature, Nature Biotechnology, Genome Biology, PLoS Genetics and PLoS Pathogens.
So far, the BIG platform has received funding and support from the Plant Health and Environment department of INRA, the CNRS, the CATI BBRIC and the Institut Sophia Agrobiotech.

Keywords: genomics, plant health, transcriptomics, evolution

The Bioinformatics and Biostatistics Hub: first results and challenges

Damien Mornico * [1], Marie-Agnès Dillies *

[2], Christophe Malabat *

[3]

[1] Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI) – Institut Pasteur de Paris – 25-28 Rue du Docteur Roux, 75015, Paris, France

[2] Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI) – Institut Pasteur de Paris – 25-28, rue du Docteur Roux 75724 Paris Cedex 15, France

[3] Centre de Bioinformatique, biostatistique et Biologie Intégrative (C3BI) – Institut Pasteur de Paris, Centre National de la Recherche Scientifique - CNRS – Hub de bioinformatique biostatistique Centre de Bioinformatique, biostatistique et Biologie Intégrative Institut Pasteur 25-28 rue du Docteur Roux, 75724 Paris CEDEX 15, France

The Bioinformatics and Biostatistics Hub was created in 2015 as part of the Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI) of the Institut Pasteur. The missions of the Hub are manifold: from quick answers at open desk sessions to short- and long-term collaborations with research units and platforms of the Parisian campus, tool and methodological development and teaching. The Hub is composed of 50 research engineers, among whom 40 have been recruited within four recruitment campaigns, between 2014 and 2017. Their expertise covers most areas of bioinformatics and biostatistics, such as algorithm, software and web development, sequencing data analysis (DNA, transcriptomics, epigenomics, metagenomics, ...), phylogenetics, statistics and statistical modeling. The Hub's model is innovative and has been designed to be adaptive to the needs of the research teams of the Institut Pasteur. More than 240 projects have been submitted to the Hub since early 2015. Project submission and management are carried out using an efficient and internally developed web application. The Hub's members can either be detached in research units or platforms for several years, embedded in research units for a long-term collaboration (from three months to two years), or integrated into the Hub core where they are in charge of short collaborations. They can also share their time between their host team and the other Hub members in the open spaces. Hub members can move from one position to another, depending on their projects and their individual needs and desire. The Hub core is structured in six open expert groups (algorithms and software development, web integration, functional genomics, transcriptomics and epigenomics, statistics, and phylogenetics) that are in charge of the projects and scientific animation associated with their area of expertise. Every Hub member devotes at least 20% of his/her time in activities that benefit to the community, such as weekly open desks, training, software and web development, or scientific animation. More than 150 hours of lectures and 450 hours of practical courses were

*Speaker

provided during the 2017-2018 academic year. This poster will provide an overview of the Hub, its missions, organization, and main results after three years of existence.

Keywords: platform, bioinformatics, biostatistics, integrative biology, NGS data analysis, web integration, development, phylogenetics

The CRCM Integrative Bioinformatics (Cibi) platform, an offer of service in latest bioinformatics technologies for large scale data analysis in biology

Quentin Da Costa [1], Samuel Granjeaud , Ghislain Bidaut * [1]

[1] Centre de Recherche en Cancérologie de Marseille (CRCM) – Aix Marseille Université : UM105, Institut Paoli-Calmettes : UMR7258, Institut National de la Santé et de la Recherche Médicale : U1068, Centre National de la Recherche Scientifique : UMR7258 – 27 bd Leï Roure, BP 30005913273 Marseille Cedex 09, France

The CRCM Integrative bioinformatics platform is composed of 3 permanent engineers, and fixed-term contracts engineers as well as several students.
Our goal is to offer a technological and multidisciplinary expertise to support interdisciplinary work in research project featuring data analysis in various biological fields related to cancer. We have technological development in various data analysis intensive areas, including Next Generation Sequencing (NGS), systems biology and gene network analysis, and multiparametric flow cytometry analysis.

Our activity is structured in two parts, research and service. For both activities, we have developed several research pipelines for data processing.

For service, we perform routine analysis in Next Generation Sequencing analysis. We have set up pipelines for RNA-Seq, Chip-Seq and variant analysis for diagnostics using latest R/Bioconductor packages. We perform standard microarray analysis (DNA and ChIP) coupled with Gene Ontology and pathway enrichment (GSEA). We also perform metabolomics and other mass spectrometry analysis. We are currently implementing all our pipelines in Mobyle and Galaxy in order to make them available to researchers. We have privileged access to computational resources of the CRCM mesocentre managed by the DISC platform.

To enforce Quality Assurance (QA), we manage all projects with our Redmine instance. All research results are giver under the form of automatically generated reports with advanced visualization, generated with R/Bioconductor. We are also using latest R technologies (Shiny) to build advanced GUIs for our pipelines.

In the research side, we routinely perform gene network analysis for systems biology. We have developed interactome analysis technologies with various research teams inside and outside CRCM. We have developed ITI (Interactome-Transcriptome Integration) a gene-expression analysis pipeline that allows detecting gene networks deregulated in cancer (Garcia et al., 2012) (Garcia et al., 2014a). We developed a version for copy-number analysis (Copy-Number Variation-

[*] Speaker

ITI, CITI, Garcia et al., 2014b), that we used to identify drivers genes and regulated functional modules in breast cancer molecular subtypes. We are now using this methodology for routine proteomics (Bonacci et al., 2014).

We recently developed a regulation analysis network for siRNA screening applications (Rioualen et al., 2017). Gene expression time series analysis is currently under finalization (data not yet published).

A framework for flow cytometry analysis was implemented and used through several collaborations (Gondois-Rey et al., 2012) (Gondois-Rey et al., 2016).

We are participating to several national and international research programs in oncology, with collaborators from various institutes (Institut National du Cancer, ITMO Cancer, H2020). On educational side, we are currently participating to in house educational program for biologists and have given several presentations from variant analysis to practical use of R for bioinformatics analysis. Platform members are experts on several university programs (University of Turin Advanced Studies Doctorate Programs, and Aix Marseille Université Master's program in Oncology).

In conclusion, we have developed an expertise in bioinformatics and data analysis, and are proposing our services for collaborations for either routine and research projects.

References

Bonacci, T., Audebert, S., Camoin, L., Baudelet, E., Bidaut, G., Garcia, M., Witzel, I.-I., Perkins, N.D., Borg, J.-P., Iovanna, J.-L., et al. (2014). Identification of new mechanisms of cellular response to chemotherapy by tracking changes in post-translational modifications by ubiquitin and ubiquitin-like proteins. J. Proteome Res. *13*, 2478–2494.

Garcia, M., Millat-Carus, R., Bertucci, F., Finetti, P., Birnbaum, D., and Bidaut, G. (2012). Interactome-transcriptome integration for predicting distant metastasis in breast cancer. Bioinforma. Oxf. Engl. *28*, 672–678.

Garcia, M., Finetti, P., Bertucci, F., Birnbaum, D., and Bidaut, G. (2014a). Detection of driver protein complexes in breast cancer metastasis by large-scale transcriptome-interactome integration. Methods Mol. Biol. Clifton NJ *1101*, 67–85.

Garcia, M., Millat-Carus, R., Bertucci, F., Finetti, P., Guille, A., Adela´ide, J., Bekhouche, I., Sabatier, R., Chaffanet, M., Birnbaum, D., et al. (2014b). CNV-Interactome-Transcriptome Integration to detect driver genes in cancerology. In Microarray Image and Data Analysis: Theory and Practice, (Luis Rueda), pp. 331–338.

Gondois-Rey, F., Granjeaud, S., Kieu, S.L.T., Herrera, D., Hirsch, I., and Olive, D. (2012). Multiparametric cytometry for exploration of complex cellular dynamics. Cytom. Part J. Int. Soc. Anal. Cytol. *81*, 332–342.

Gondois-Rey, F., Granjeaud, S., Rouillier, P., Rioualen, C., Bidaut, G., and Olive, D. (2016). Multi-parametric cytometry from a complex cellular sample: Improvements and limits of manual versus computational-based interactive analyses. Cytom. Part J. Int. Soc. Anal. Cytol. *89*, 480–490.

Rioualen, C., Da Costa, Q., Chetrit, B., Charafe-Jauffret, E., Ginestier, C., and Bidaut, G. (2017). HTS-Net: An integrated regulome-interactome approach for establishing network regulation models in high-throughput screenings. PloS One *12*, e0185400.

Keywords: Bioinformatics platform, service platform, data analysis, genomics and functional analysis, transcriptomics, Next Generation Sequencing, proteomics, interactome, gene networks, NGS diagnostics

MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic and metabolic comparative analysis

David Vallenet* [1], Alexandra Calteau [1], Stephane Cruveiller [1], Mathieu Dubois †‡ [1], Aurélie Lajus [1], David Roche [1], Zoe Rouy [1], Mylène Beuvin [1], Celine Chevalier †

[1], Mathieu Gachet [1], Guillaume Gautreau †

[1], Jordan Langlois [1], Rémi Planel †

[1], Johan Rollin †

[1], Valentin Sabatet [1], Claudine Médigue§ [1]

[1] Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme (LABGeM) – CEA, Genoscope : DRF/IBFJ/GEN, CNRS : UMR8030, Université d'Evry - Université Paris-Saclay – 91057 Evry, France

Introduction

Large-scale genome sequencing and the increasingly massive use of high-throughput approaches produce a vast amount of new information that completely transforms our understanding of thousands of microbial species. However, despite the development of powerful bioinformatics approaches, full interpretation of the content of these genomes remains a difficult task. To address this challenge, we have developed the MicroScope platform [1,2], which is a software environment for management, annotation, comparative analysis and visualization of microbial genomes (https://www.genoscope.cns.fr/agc/microscope). Published for the first time in 2006 [3], the platform has been under continuous development within the LABGeM group at Genoscope and provides analysis for complete and ongoing genome projects together with post-genomic experiments (i.e. transcriptomics, re-sequencing of evolved strains, mutant collections) allowing users to improve the understanding of gene functions.

[*] Corresponding author: vallenet@genoscope.cns.fr

[†] Speaker

[‡] Corresponding author: mdubois@genoscope.cns.fr

[§] Corresponding author: cmedigue@genoscope.cns.fr

Who is using MicroScope and for what purposes?

MicroScope serves different use cases in bioinformatics:

> it supports the integration of newly sequenced or already available prokaryotic genomes through the offer of a free-of-charge service to the scientific community
>
> it performs computational inferences including prediction of gene function, metabolic pathways, resistome and virulome
>
> it provides tools for comparative genomic and metabolic analyses and visualization
>
> it supports collaborative expert annotation processes through the use of specific curation tools and graphical interfaces.

To date, MicroScope contains data for _˜10,000 microbial genomes, which are manually curated and analyzed by microbiologists (> 3,800 personal accounts in April 2018). The platform enables collaborative work and improves community-based curation efforts in a rich comparative genomic context. Indeed, gene context approaches often complement the classical homology-based gene annotation for assigning function to novel proteins. MicroScope has been used to perform a complete expert annotation of several reference species such as Escherichia coli [3,4], Bacillus subtilis 128 [5,6], Pseudomonas putida KT2440 [7], Acinetobacter baylyi APD1 [8]. In addition, important pathogens and environmental species have also been extensively analyzed.

What's next?

Here, we present an overview of the MicroScope analysis pipelines and illustrate the use of several new functionalities which concern:

> the evaluation of genome completion and contamination with CheckM software [9] and the computation of genome clusters (i.e. species group) using genomic distances that are estimated using Mash software [10]
>
> comparative genomics using a graph approach to model pangenomes (PPanGGOLiN method, https://github.com/ggautreau/PPanGGOLiN) and classify gene families into three partitions (i.e. persistent, shell and cloud)
>
> the detection of regions of genomic plasticity and the analysis of their gene content (i.e. virulence and antimicrobial resistance genes, secretion systems, integrons and secondary metabolite biosynthesis gene clusters)
>
> rule-based annotation using rules to predict functions (UniRule system) and the GROOLS expert system (https://github.com/grools) that assists biologists in the curation of genes involved in metabolic pathways by highlighting uncertainties and inconsistencies [11].

1. Vallenet D, Calteau A, Cruveiller S, Gachet M, Lajus A, Josso A, et al. MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. Nucleic Acids Res. Oxford University Press; 2016;45: D517–D528.

2. Médigue C, Calteau A, Cruveiller S, Gachet M, Gautreau G, Josso A, et al. MicroScope-an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data. Brief Bioinform. 2017;

3. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, et al. MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res. Oxford University Press; 2006;34: 53–65.

4. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. Public Library of Science; 2009;5: e1000344.

5. Borriss R, Danchin A, Harwood CR, Médigue C, Rocha EPC, Sekowska A, et al. Bacillus subtilis, the model Gram-positive bacterium: 20 years of annotation refinement. Microb Biotechnol. 2018;11: 3–17.

6. Belda E, Sekowska A, Le Fèvre F, Morgat A, Mornico D, Ouzounis C, et al. An updated metabolic view of the Bacillus subtilis 168 genome. Microbiology. Microbiology Society; 2013;159: 757–770.

7. Belda E, van Heck RGA, José Lopez-Sanchez M, Cruveiller S, Barbe V, Fraser C, et al. The revisited genome of Pseudomonas putida KT2440 enlightens its value as a robust metabolic chassis. Environ Microbiol. 2016;18: 3403–3424.

8. Barbe V, Vallenet D, Fonknechten N, Kreimeyer A, Oztas S, Labarre L, et al. Unique features revealed by the genome sequence of Acinetobacter sp. ADP1, a versatile and naturally transformation competent bacterium. Nucleic Acids Res. 2004;32: 5766–5779.

9. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25: 1043–1055.

10. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17. doi:10.1186/s13059-016-0997-x

11. Mercier J, Josso A, Médigue C, Vallenet D. GROOLS: reactive graph reasoning for genome annotation through biological processes. BMC Bioinformatics. 2018;19: 132.

Keywords: microbial genomics, functional annotation, comparative genomics, metabolism

Montpellier GenomiX (MGX) : next-generation sequencing and data analysis service and expertise

Stéphanie Rialle * [1], Emeric Dubois , Marine Pratlong

[1] Montpellier GenomiX (MGX) – CNRS : UMS3426 – France

The MGX (Montpellier GenomiX) is an ISO 9001:2015 certified facility which offers, since 2009, next-generation sequencing services, as well as bioinfomatics and biostatistics analysis of the produced data. The facility is accessible to both academic and industry/biotech scientists. Our expertise is drawing on many years of experience, as much in molecular biolology as in bioinformatics.

We propose a wide range of applications, including whole-genome sequencing, exome and targeted sequencing, RNA-seq, small RNA-seq, epigenetics (ChIP-seq, HiC, Whole Genome and Reduced Representation Bisulfite Sequencing, ...), population genomics with RAD-seq, etc. We cover all steps from library construction to bioinformatics analysis. Bioinformatics analyses include quality control, alignment of sequences to genome or transcriptome, statistical and functional analyses. A typical project starts with a launch meeting to define the aims of the experiment, the experimental design, and the analysis tools to be used. Throughout the project, a project management web application provides an easy and flexible way to store and retrieve information, and to communicate with customers.

We lately made the acquisition of the Chromium machine from 10x Genomics, and thus propose single cell gene expression, as well as linked reads analysis. The Chromium is based on a droplet-based approach, which allows the characterization of hundreds to millions of cells in a single experiment. The 3' mRNA quantification of gene expression can lead to the identification of cell population in heterogenous or complex samples. The analysis can be done using the Cell Ranger solution available from 10x Genomics. We are also currently evaluating other approaches to improve the results concerning normalization, dimension reduction, clustering, statistical analysis and data visualization. The linked reads solution enables *de novo* diploid genome assembly, reconctruction of long range haplotypes, and can improve complex strucutal variant detection.

Besides sequencing on an Illumina machine, we also offer the possibility to sequence on MinION (Oxford Nanopore Technologies), which produces long reads from DNA or RNA samples. In a first step, we particularly focused on direct RNA sequencing using the nanopore technology which allows to directly sequence complete RNA molecules (without the cDNA synthesis step). After several weeks of tests and tuning of the protocol, we are now able to produce routinely nearly 1 million reads with a N50 of around 1 300 bp. Our sequencing runs from human mRNA allowed to align 76 % of reads using GMAP and are currently being analysed to study the mechanism of alternative splicing and intron retention. More recently, we performed our first

*Speaker

tests of the DNA sequencing protocole using the nanopore technology. We are currently testing a panel of dedicated assemblers in order to settle a robust pipeline of genome assembly using nanopore long reads.

Keywords: next, generation sequencing, single cell, long reads, epigenetics, RNA, seq, genomics, statistical analysis, MinION, 10x Genomics

Community support for Systems Ecology

Marie Chevallier * [1,2]

[1] Ecosystèmes, biodiversité, évolution [Rennes] (ECOBIO) – Universite de Rennes 1, INEE, Observatoire des Sciences de l'Univers de Rennes, Centre National de la Recherche Scientifique : UMR6553 – Bâtiment 14 - Université de Rennes 1 - Campus de Beaulieu - CS 74205 - 35042 Rennes Cedex - France, France

[2] Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) – Universite de Rennes 1, Institut National des Sciences Appliquées - Rennes, Université de Bretagne Sud, École normale supérieure - Rennes, Institut National de Recherche en Informatique et en Automatique, CentraleSupélec, Centre National de la Recherche Scientifique : UMR6074, IMT Atlantique Bretagne-Pays de la Loire – Avenue du général LeclercCampus de Beaulieu 35042 RENNES CEDEX, France

EcoSyst is the Biogenouest Federator Project. It aims to support the emergence of Systems Ecology in Western France. Thanks to the strengths and skills present, EcoSyst is the incubator of new ideas and interdisciplinary projects. By inter-regional animation (Brittany and Pays de la Loire) EcoSyst highlights local expertise and allows the appropriation of issues related to complex ecosystems. This project aims to support scientific and technical innovation, stimulate the emergence of new research projects and focus on proximity to the Western France actors to better respond to their needs. Under the impetus of this dynamic, the scientific community benefits from access to a set of methods to develop the identification of emerging properties at the community-scale of biological organisms and more broadly at the ecosystem level, by integrating various layers of hierarchical organizations.
The Ecosyst project enabled to create a network of researchers and engineers concerned by the issues of systems ecology. This domain is based on the theory of complex systems that analyzes the emergence of global properties resulting from the interaction of system components. By adapting these theories to ecosystems, the challenge here is to identify these emergent properties at the organisms community level. The necessary methods to identify these emergent properties include: sequence bioanalysis (identification of species in environmental samples, microbiota in the broad sense) ; bio-statistics (identification of correlations) ; network theory (interpretation of data by co-occurrence or metabolic graphs to abstract the exchanges of biological compounds structuring the microbial communities) ; the theory of evolution (natural selection, Red Queen hypothesis, Black Queen hypothesis) and the ecological modeling (numerical models allowing the mechanistic simulation of communities or different evolutionary pressure hypotheses on a system). The vitality of this new area of research can be seen in the important work published recently (eg Zamorrodi & Segré, Nat Comm 2017, Carrier & Reitzel, Nat Comm 2018). In particular, these works undertake key reflections on the predictability of biological or microbiological phenomena and the assessment of the importance of uncertainty in models (Delahaye et al, MSystems 2018).

The choice to focus on microbiota is related to the fact that the genome of microorganisms

*Speaker

is generally less complex than that of multicellular eukaryotes. In addition, microbiota associated with macro-organisms represent a very important research issue (read for example in PLoS Biology 'How the microbiome challenges our concept of self' (Rees et al., 2018)).
In its first phase (2016-2018), the EcoSyst project, aiming to bring out systems ecology in West France, has confirmed that the forces and skills are present, even if they are spread among different research structures. Having highlighted the expertise in ecology, environment, modeling, and their applications on species of interest in agronomy, sea and health, the project has identified a complete network of actors and key structures on the Brittany - Pays de la Loire inter-region (19 structures involved in Angers, Brest, Nantes, Rennes, Roscoff and Saint Nazaire). It has also fostered the emergence of new ideas and new multidisciplinary projects.

This work of animation and analysis of the need, closer to the local scientific community, made it possible to highlight the main biological questions of the field. Blatantly, whatever the discipline, we find many common questions:

- Composition: Who is there? In which conditions?
- Interactions: Who interacts with whom? How?
- Prediction: How is the system evolving?
- Markers: Who does what?

And in the same way, in the study of organisms communities (composition, functions ...), a certain number of tools are commonly used, and in particular derived from Systems Biology (Frogs, SWARM, MegaRast, MetExplore...). However, we also realize the many limitations to the use of these tools. Firstly at the level of installation and maintenance of tools, in teams where there is not necessarily proximity to a computer service or in-house expertise. Also, the usability of these tools is low since they often require strong expertise (often missing in teams). In addition, the research community increasingly faces the problems of scaling up large data sets, necessarily used for the study of complex systems. This also requires a strong competence in bioinformatics to understand the issues and know how to optimize the algorithms to extract knowledge. Moreover, as systems ecology is an emerging field, there is no well-defined study protocol or methodology to study a complete system. Therefore, there is no guide or standard process and it is necessary for a researcher to be aware and compare existing tools, to be able to install them and connect them by hand, to be able in the end to analyze results... The very complexity of these systems often requires the integration of various methods to cross the scales of study.

In an attempt to address these limitations, a proof of concept has been developed: a pipeline for modeling interactions within complex organisms communities. This approach by network modeling of OTUs, crossed with methods of metabolic meta-networks analysis, made it possible to suggest interesting hypotheses on the key-interactions within a complex system, to be validated in vitro.

However, this first component has highlighted the lack of skill by the community on existing methodologies and tools for systems ecology research. This may be explained by the necessarily multidisciplinary nature of this research area, where it has become very complicated for teams to have in their workforce the set of skills involved. A number of tools and methods are present on support platforms (such as ABIMS, BIRD, GenOuest), but this offer remains incomplete and unknown. In addition, the plurality of platforms scrambles the service offering and loses users.

The proposed prototype pipeline, characterizing a first service integration effort, will be made available to the community in summer 2018. It will be distributed in the form of various Dockers containers deployed via the existing support platforms and will integrate different modules: annotation of genomes and metagenomes (ABIMS), diversity analysis (BIRD), co-occurences study (BIRD, genomic platform related to GenOuest), metabolic analysis (GenOuest, BIRD), environmental modeling (BIRD). The scientific community network will be structured via the CesGo e-science environment. The ultimate goal is to implement a complete ecosystem modeling solution, based on multi-dimensional data (genomic sequences, meta-genomics, transcriptomic data and meta-transcriptomics), and distributed on infrastructures (servers, platforms).

Keywords: systems ecology, community, network, microbiota, tools, platform, pipeline

South Green, une plateforme de bioinformatique tournée vers l'agriculture dans les pays du sud

Sébastien Ravel *† [1,2], Stéphanie Bocs [2,3], Louise Brousseau [2,4], Frédéric De Lamotte [2,3], Alexis Dereeper [2,5], Gaëtan Droc [2,3], Jean-François Dufayard [2,3], Valentin Guignon [2,6], Chantal Hamelin [2,3], Pierre Larmande [2,4], Frédéric Mahé [2,7], Guillaume Martin [2,3], Julie Orjuela-Bouniol *

[2,4,5], Bertrand Pitollat [2,3], Mathieu Rouard [2,6], Manuel Ruiz [2,3], François Sabot *

[2,4], Gautier Sarah [2,3], Guilhem Sempéré [2,8], Maryline Summo [2,3], Ndomassi Tando [2,4], Christine Tranchant-Dubreuil [2,4]

[1] Biologie et génétique des interactions plantes-parasites pour la protection intégrée (BGPI) – Institut national de la recherche agronomique (INRA) : UR0385, Centre de coopération internationale en recherche agronomique pour le développement [CIRAD] : UMR54 – Campus International de Baillarguet - TA 41 / K - 34398 Montpellier Cedex 05, France
[2] South Green Bioinformatics Platform (SG) – Bioversity, CIRAD : UMRAGAP / BGPI / LSTM, Institut de recherche pour le développement [IRD] : UMRDIADE/ IPME – Montpellier, France
[3] CIRAD UMR AGAP (AGAP) – Institut national de la recherche agronomique (INRA) : UMR1334 – TA A-108/03-Avenue Agropolis, 34398 Montpellier Cedex 5, France
[4] UMR DIADE IRD/UM (DIADE) – Université de Montpellier, Institut de Recherche pour le Développement – Centre IRD de Montpellier 911 av Agropolis BP 604501 34394 Montpellier cedex 5, France
[5] IRD IPME (IPME) – Institut de recherche pour le développement [IRD] – Avenue Agropolis, 34398 Montpellier Cedex 5, France
[6] Bioversity International – Consultative Group on International Agricultural Research [CGIAR] – Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France
[7] LSTM, Univ Montpellier, CIRAD, IRD, INRA, Montpellier SupAgro – LSTM – Montpellier, France
[8] UMR Intertryp - CIRAD - IRD – Institut de recherche pour le développement [IRD], CIRAD – Avenue Agropolis - 34398 Montpellier Cedex 5, France

South Green (www.southgreen.fr) est une plateforme de bioinformatique dédiée à la génétique et la génomique des plantes tropicales et méditerranéennes d'intérêt agronomique et de leurs pathogènes. Elle fédère un réseau de bioinformaticiens appartenant à différentes unités et instituts de Montpellier (Bioversity International, CIRAD, INRA et IRD) soit environ une vingtaine de personnes en interaction avec les équipes de recherches, avec une expertise multidisciplinaire

*Speaker

†Corresponding author: sebastien.ravel@cirad.fr

allant de l'intégration de données et de connaissance au développement de logiciels en bioinformatique, à l'analyse de données de séquençage (détection de polymorphismes et variants structuraux, pangénomique, métagénomique, analyse différentielle de données RNAseq) et le calcul haute performance.

Le plateforme South Green a pour objectifs de :

Promouvoir des outils originaux issus de la recherche méthodologique.

Promouvoir l'interopérabilité des applications développées au sein du réseau.

Centraliser l'ensemble des logiciels et systèmes d'information développées au sein d'un portail Web unique (http://www.southgreen.fr)

Promouvoir les échanges et les développements collaboratifs

Proposer des formations en bioinformatique, bioanalyse de données et à l'utilisation de clusters de calcul

Promouvoir la démarche qualité au sein du réseau.

Proposer un support pour le calcul à haute performance

La plateforme assure le développement de systèmes d'informations et d'outils innovants, nécessaires aux projets scientifiques, réalisés au sein de la plateforme et en lien avec l'analyse des données produites par les technologies de séquençage à haut débit (annotation des génomes et de transcriptomes, phylogénie, génotypage) tels que GreenPhylDB , SNiPlay, Gigwa ou AgroLD. Elle propose également des pipelines d'analyses de données de séquençage au travers de deux gestionnaires de workflows : Galaxy et TOGGLe. Enfin, impliquée dans plusieurs projets de séquençage international, elle possède une forte expertise en développement de "genome hub" qu'elle a déployée sur de nombreuses plantes (bananier, caféier, manioc, cacaoyer) au niveau duquel sont disponibles de nombreuses applications utiles pour l'étude de ces génomes.

La plateforme assure aussi des formations spécialisées en bioinformatique au niveau national et international (analyse des données de séquençage haut débit, Galaxy) et informatique (logiciel R, Perl, Linux). Les ressources sont disponible sur le site https://southgreenplatform.github.io/trainings/. South Green s'inscrit dans le réseau des plateformes de l'Institut Français de Bioinformatique (IFB) et fait partie du réseau Renabi (Réseau national des plates-formes bioinformatiques).

Reference :

South Green collaborators. The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics (2016) Curr. Plant Biology dx.doi.org:10.1016/j.cpb.2016.12.002

Liens:

www.southgreen.fr

https://github.com/SouthGreenPlatform

Keywords: plateform, workflow manager, genome hub, data interoperability, data mining, tools, training, HPC

The Systems Biology Graphical Notation: a standardised representation of biological maps

Vasundra Touré *† [1], Alexander Mazein [2], Adrien Rougny [3], Andreas Dräger [4], Ugur Dogrusoz [5], Augustin Luna [6], Nicolas Le Novère [7]

[1] Department of Biology, Norwegian University of Science and Technology [Trondheim] (NTNU) – NO-7491 Trondheim, Norway

[2] European Institute for Systems Biology and Medicine – CIRI, Inserm, U1111, Université Claude Bernard Lyon 1, CNRS, UMR5308, École Normale Supérieure de Lyon, Univ Lyon, F-69007, Lyon, France – 50 Avenue Tony Garnier, 69007 Lyon, France

[3] Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology – Aomi, Tokyo 135-0064, Japan

[4] Applied Bioinformatics Group, Center for Bioinformatics Tübingen (ZBIT), University of Tübingen – Sand 14 C320 72076 Tübingen, Germany

[5] Computer Engineering Department, Bilkent University – Bilkent University 06800 Bilkent, Ankara TURKEY, Turkey

[6] cBio Center, Dana-Farber Cancer Institute, Boston, MA; Department of Cell Biology, Harvard Medical School – Boston, MA 02215, United States

[7] The Babraham Institute – Babraham Hall, Babraham, Cambridgeshire CB22 3AT, United Kingdom

Background: Visualization of biological processes plays an essential role in life science research. Over time, diverse forms of diagrammatic representations, akin to circuit diagrams, have evolved without well-defined semantics potentially leading to ambiguous network interpretations and difficult programmatic processing.

Results: The Systems Biology Graphical Notation (SBGN) is a standard developed to reduce ambiguity in the visual representation of biomolecular networks. It provides specific sets of well-defined symbols for various types of biological concepts. SBGN comprises three complementary languages: Process Description (PD), Entity Relationship (ER), and Activity Flow (AF). SBGN PD is based on reactions and is well-suited for detailed sequential biochemical mechanisms, for instance, to represent metabolic pathways. SBGN AF shows cascades of influences between the activities carried by biomolecular entities (e.g., stimulation, inhibition) and is particularly useful when the precise molecular mechanisms are unknown or do not need to be shown, for instance, to represent signalling pathways and regulatory networks. SBGN ER represents independent interactions between features of biological entities, which avoids combinatorial explosions of represented biological states and interactions. The XML-based SBGN Markup Language (SBGN-ML) facilitates convenient storage and exchange of SBGN maps, supported by the library libSBGN.

Discussion: The SBGN project is an ongoing open community-driven effort coordinated and

*Speaker

†Corresponding author: vasundra.toure@ntnu.no

maintained by an elected international editorial board. Annual workshops, GitHub and mailing lists are used as leading discussion platforms. Major research projects, such as the Virtual Metabolic Human, and pathway databases such as Reactome and WikiPathways display their maps following the SBGN guidelines. Furthermore, a wide range of tools supports SBGN. SBGN regularly offers student coding events through the Google Summer of Code program.

Availability: All documents and source code are freely available at http://sbgn.org and https://github.com/sbgn
Contributions are welcome.

Contact: sbgn-discuss@googlegroups.com

References

Le Novère, Nicolas, Michael Hucka, Huaiyu Mi, Stuart Moodie, Falk Schreiber, Anatoly Sorokin, Emek Demir, et al. 2009. ”The Systems Biology Graphical Notation.” Nature Biotechnology 27 (8): 735–41. doi:10.1038/nbt.1558.

Vasundra Touré, Nicolas Le Novère, Dagmar Waltemath and Olaf Wolkenhauer. 2018. ”Quick tips for creating effective and impactful biological pathways using the Systems Biology Graphical Notation”. PLoS Comput Biol 14(2): e1005740. doi:10.1371/journal.pcbi.1005740.

Stuart Moodie, Nicolas Le Novère, Emek Demir, Huaiyu Mi, Alice Villéger. 2010 ”Systems Biology Graphical Notation: Process Description language Level 1 Version 1.3.” doi:10.2390/biecoll-jib-2015-263.

Anatoly Sorokin, Nicolas Le Novère, Augustin Luna, Tobias Czauderna, Emek Demir, Robin Haw, Huaiyu Mi, et al. 2015. ”Systems Biology Graphical Notation: Entity Relationship language Level 1, Version 2.” doi:10.2390/biecoll-jib-2015-264.

Huaiyu Mi, Falk Schreiber, Stuart Moodie, Tobias Czauderna, Emek Demir, Robin Haw, Augustin Luna, et al. 2015. ”Systems Biology Graphical Notation: Activity Flow language Level 1, Version 1.2.” doi:10.2390/biecoll-jib-2015-265.

Martijn van Iersel, Alice Villéger, Tobias Czauderna, Sarah Boyd, Frank Bergmann, Augustin Luna, Emek Demir et al. 2012. ”Software support for SBGN maps: SBGN-ML and LibSBGN.” Bioinformatics, 28(15):2016-2021. doi:10.1093/bioinformatics/bts270.

Keywords: SBGN, circuit diagram, biological network, visualisation, systems biology

The StatOmique group is ten years old

Marie-Agnès Dillies *† 1, Julie Aubert‡ 2, Christelle Hennequet-Antier§ 3

[1] Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI) – Institut Pasteur de Paris – 25-28, rue du Docteur Roux 75724 Paris Cedex 15, France
[2] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France (UMR MIA-Paris) – Institut National de la Recherche Agronomique - INRA, AgroParisTech – AgroParisTech 16 rue Claude Bernard 75 005 PARIS, France
[3] Unité de Recherches Avicoles (URA) – Institut national de la recherche agronomique (INRA) : UR0083 – 37380 Nouzilly, France

The StatOmique group was created in 2008 with about 15 members. At that time, the main aim was to help isolated practitioners involved in expression data analysis to exchange about methods and best practice. After 10 years of existence the group has evolved and gathers about 60 statisticians and bioinformaticians. It is interested in statistical and bioinformatics methods and tools developed for the analysis of high-throughput genomic data. It has invested more particularly in the field of transcriptomics, but also other fields of application such as epigenomics, metagenomics, metabolomics or statistics of genome-wide association. Omics data are characterized by "small n, large p", where the number of variables measured is always much greater than the number of individuals in the experiment. Consequently, it has been necessary in recent years to develop statistical methods adapted to new technologies and to the problem of the large dimension (dimension reduction, variable selection, regularization, etc.). In addition, the emergence of single cell and long reads technologies induces the development of new methods adapted to the biological questions to which they now allow to respond. StatOmique is also interested in multivariate statistical approaches as part of the challenge of integrating different types of data. The poster will describe the history of the group and its main achievements, as well as the work in progress and future challenges that it wishes to address.

Keywords: Statistics, Expression data, NGS, RNA, seq data normalization, multivariate analysis

*Speaker

†Corresponding author: marie-agnes.dillies@pasteur.fr

‡Corresponding author: julie.aubert@agroparistech.fr

§Corresponding author: christelle.hennequet@tours.inra.fr

3'seq-RP : A low cost high throughput digital sequencing technique to evaluate gene expression profiles

Eric Charpentier *† [1], Audrey Bihouee * ‡ [1], Dimitri Meistermann [2], Stéphanie Kilens [2], Léa Flippe [2], Solenne Dumont [1], Audrey Donnart [1], Marine Cornec [1], Laurent David [2], Richard Redon [1]

[1] unité de recherche de línstitut du thorax UMR1087 UMR6291 (ITX) – Université de Nantes, Institut National de la Santé et de la Recherche Médicale : U1087, Centre National de la Recherche Scientifique : UMR6291 – 8 quai Moncousu - BP 70721 - 44007 Nantes Cedex 1, France
[2] Centre de Recherche en Transplantation et Immunologie (CRTI) – Université de Nantes, Institut National de la Santé et de la Recherche Médicale : U1064 – CHU Nantes, 30 Bd Jean Monnet, 44093 Nantes Cedex 1, France

Introduction
RNA-seq has become the gold standard approach to evaluate genome wide transcriptome profiles. Experimental design has to balance a sufficiently high number of samples with an affordable experiment. Hence, the choice between the number of tested conditions and the number of replicates is often at the expense of the latter. The costs come from 2 steps: the throughput of the library preparation and the sequencing depth. Typically, sequencing depth is important for gene modeling and allelic expression analysis. A method improving the throughput of the library preparation and limiting the required sequencing depth would allow powerful analysis on large cohorts of samples while lowering the costs.

A new quantitative method based on molecular indexing of mRNA molecules called 3'seq-RP for "3' Sequencing RNA Profiling " has been implemented in our core facility - GenoBiRD.

This technique has several advantages: the bar-coding of the unique molecules (UMI) which allows an absolute quantification of the transcripts and its low cost (40 by samples). The samples are multiplexed on 96-well plates and sequenced on a Hiseq2500 rapid-run.

On the bioinformatics side, the standard RNAseq tools can be used with some adjustments related to UMI counting.

Applied to single cell analysis, this approach is extremely promising for characterizing the heterogeneity of cell populations, even from very low amount of material. Moreover, 3'seq-RP enables the development of a knowledge base of gene expression profiles through quantitative expression values. This will be an important resource for the unbiased comparison of profiles obtained in different physiological or pathological conditions or in different states of cell differentiation.

*Speaker
†Corresponding author: eric.charpentier@univ-nantes.fr
‡Corresponding author: audrey.bihouee@univ-nantes.fr

Protocol

3'seq-RP protocol is performed according to Ref.1. The libraries are prepared from 10ng of total RNA in 4μl. The mRNA poly(A) tails are tagged with universal adapters, well-specific barcodes and unique molecular identifiers (UMIs) during template-switching reverse transcriptase. Barcoded cDNAs from multiple samples are then pooled, amplified and tagmented using a transposon-fragmentation approach which enriches for 3ends of cDNA. A library of 350–800bp length is run on an Illumina HiSeq 2500 using a Hiseq Rapid SBS Kit v2-50 cycles (ref FC-402-4022) and a Hiseq Rapid PE Cluster Kit v2 (ref PE-402-4002).

We start the analysis by generating a sample sheet with essential informations : sample name with corresponding barcodes, sample annotation and species. Raw fastq pairs used for analysis matched the following criteria: the 16 bases of the first read correspond to 6 bases for a designed well-specific barcode and 10 bases for a unique molecular identifier (UMI). The second read (57 bases) corresponds to the captured poly(A) RNAs sequence. We perform demultiplexing of these fastq pairs according to the samplesheet to generate one single-end fastq for each of the 96 samples. These fastq files are then aligned with bwa to the reference mRNA sequences and the mitochondrial genomic sequence, both available from the UCSC download site.

DGE profiles are generated by parsing the alignment files (.bam) and counting for each sample the number of unique UMIs associated with each RefSeq genes. Reads aligned on multiple genes, containing more than one mismatch with the reference sequence or reads containing a polyA pattern are discarded. Finally, a matrix containing the expression of all genes on all samples is produced. The expression values, corresponding to the absolute abundance of mRNAs in all samples, is then ready for further gene expression analysis. DESeq2 is used to normalize expression with the DESeq function (Ref.2). Normalized counts are transformed with vst (variance stabilized transformation) function from DESeq library. Batch effects may be corrected with the limma library function "removeBatchEffect".

Results

Counting reads by RNAseq depends on the amount of total reads (library size) and the length of the gene. Conversely, the UMI barcoding in the 3'seq-RP high throughput technique allows absolute quantification of mRNA molecules that facilitates gene to gene comparison. However, transcriptome analysis by 3'seq-RP is limited to quantitative studies. Indeed, only the 3 'end of the genes is captured and sequenced, excluding isoform reconstruction, novel gene discovery or SNP studies.

Thirteen Hiseq runs have already been performed on our core facility on human and rat samples. Different types of samples were tested: cell culture, primary cells, cell sorting, biopsy.

On average, about 300 million reads are generated by a Hiseq rapid run multiplexing 96 samples. From a total of 27k human genes referenced in RefSeq, a maximum of about 16k expressed genes are detected per sample. This plateau is reached with 5 million reads. 1.5 million raw reads are enough to detect 10k genes.

Despite the absolute quantification avantage, we faced some classical biais like batch effect between runs which can be adjusted by standard tools. Moreover, we faced some more specific troubleshooting like ”neighbourhood” effect. Indeed, very highly expressed genes (like ALB) in a single sample can be detected at a very low level in all wells of a run while these genes should not be expressed in other samples. We suspect an incorrect barcode assignment.

To automate the analysis of the 3’seq-RP, we developed a Snakemake pipeline based on one developed at the Broad Technology Labs (Ref.3)

This pipeline is part of our registry and based on FAIR practices taking advantage of virtual environments (conda, docker) and continuous integration (jenkins). This best practices allow pipelines to be deployed in multiple environments for reproducibility and scalability issues. The pipeline generates a HTML report based on a json data description, and JavaScript/Boostrap/jinja/python technologies. This report, intended for end-users, displays project summary, raw and processed data quality controls and differential analysis results (PCA, differential expressed genes, sample clustering...)

Three projects in immunology domain using 3’eRP-seq have been published, two about transplantation (Ref.6,7), and one about CD8+ T cells in multiple sclerosis (Ref.8).

Lastly, the technique has been validated in a context of single cell and pluripotent stem cells (Ref.5). The transcriptome analysis allowed the validation of a protocol that can reprogram somatic cells directly to a state resembling the human preimplantation epiblast, without an intermediate passage in primed media.

Conclusions

The digital high throughput 3’ end RNA profiling is now well implemented on our genomics and bioinformatics core facilities. It seems to be a promising technique to quantify gene expression in cost effective way (20 times cheaper than RNAseq), with many biological replicates. It allows to easily compare transcriptomic profiles of different cellular types and species. We can consider building a data warehouse of gene expression patterns and modeling the expression levels according to gene functions. This protocol is also easily adaptable to single-cell transcriptome analysis by adding a new cell index, like used in Ref 5.

References

1. Magali Soumillon, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, Tarjei S Mikkelsen. Characterization of directed differentiation by high-throughput single-cell RNA-Seq.doi: https://doi.org/10.1101/003236

2. Love MI, Huber W and Anders S (2014).Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15, pp. 550. doi: 10.1186/s13059-014-0550-8.

3. https://www.broadinstitute.org/broad-technology-labs/high-throughput-eukaryote-3-dge-library-construction

4. D. Cacchiarelli, C. Trapnell, M.J. Ziller, M. Soumillon, M. Cesana, R. Karnik, et al. Integrative analyses of human reprogramming reveal dynamic nature of induced pluripotency. Cell, 162 (2) (2015 Jul 16), pp. 412-424

5. Kilens S, Meistermann D, Moreno D, Chariau C, Gaignerie A, Reignier A, Lelièvre Y, Casanova M, Vallot C, Nedellec S, Flippe L, Firmin J, Song J, Charpentier E, Lammers J,

Donnart A, Marec N, Deb W, Bihouée A, Le Caignec C, Pecqueur C, Redon R, Barrière P, Bourdon J, Pasque V, Soumillon M, Mikkelsen TS, Rougeulle C, Fréour T, David L and The Milieu Intérieur Consortium. Parallel derivation of isogenic human primed and naive induced pluripotent stem cells. Nat Commun. 2018 Jan 24;9(1):360 — https://doi.org/10.1038/s41467-017-02107-w

6. Picarda E, Bézie S, Boucault L, Autrusseau E, Kilens S, Meistermann D, Martinet B, Daguin V, Donnart A, Charpentier E, David L, Anegon I and Guillonneau C. Transient antibody targeting of CD45RC induces transplant tolerance and potent antigen-specific regulatory T cells. JCI Insight 2017 — http://dx.doi.org/10.1172/jci.insight.90088

7. S. Bézie, D. Meistermann, L. Boucault, S. Kilens, J. Zoppi, E. Autrusseau, A. Donnart, V. Nerrière-Daguin, F. Bellier-Waast, E. Charpentier, F. Duteille, L. David, I. Anegon and C. Guillonneau. Ex vivo expanded human non-cytotoxic CD8+CD45RClow/- tregs efficiently delay skin graft rejection and GVHD in humanized mice. Front Immunol, 24 January 2018 — https://doi.org/10.3389/fimmu.2017.02014

8. Nicol B, Salou M, Vogel I, Garcia A, Dugast E, Morille J, Kilens S, Charpentier E, Donnart A, Nedellec S, Jacq-Foucher M, Le Frère F, Wiertlewski S, Bourreille A, Brouard S, Michel L, David L, Gourraud PA, Degauque N, Nicot AB, Berthelot L, Laplaud DA. An intermediate level of CD161 expression defines a novel activated, inflammatory, and pathogenic subset of CD8+ T cells involved in multiple sclerosis. J Autoimmun, Oct 2017 — https://doi.org/10.1016/j.jaut.2017.10.005

Keywords: transcriptomic, DGE, umi, gene expression, transcriptomique

Third Generation Sequencing Technologies to Decipher Genomic Structures of Recombinant-prone Viruses

Mathias Vandenbogaert *† [1], Charlotte Balière [1], Aurélia Kwasiborski [1], Véronique Hourdel [1], Laure Diancourt [1], Labib Bakkali Kassimi [2], Sandra Blaise-Boisseau [2], Christophe Batejat [1], Stephan Zientara [2], Jean-Claude Manuguerra [1], Valérie Caro [1]

[1] Environment and Infectious Risks Research and Expertise Unit (ERI-CIBU) – Institut Pasteur – 25-28 rue du docteur Roux, F-75724 Paris cedex 15, France

[2] Animal Health Laboratory – Anses – 14 rue Pierre et Marie Curie, 94700 Maisons Alfort, France

INTRODUCTION/CONTEXT

The development of the ”third-generation sequencing” platforms, such as the Pacific Biosciences PacBio sequencing system and more recently the Oxford Nanopore MinION device, have yet to be exploited, and have generated particular interest within the scientific community. These methodologies open up new possibilities, such as providing the capability for minimal library preparation and long reads (up to 10 kilobases), thus enabling true linkage to be established between variants within single genomes, and resolving assembly issues that often give incorrect genomic organization.

These long-read sequencing platforms especially facilitate the analysis of viral genome structure, including recombination events, which are generally difficult to ascertain using second-generation platforms such as Illumina and Ion Torrent. In fact, current second-generation sequencing technologies have played a driving role to address questions relating to viral genome organization, epidemiology, and investigations of outbreaks by characterizing both partial- (such structural proteins) and whole-genome sequencing (WGS). In the specific case of recombinant-prone viruses, e.g. members of *Picornaviridae* family, second-generation sequencing technologies have often unveiled the limits of the approach, notably when determining precise viral genomic reconstruction and recombination hotspots.

Foot-and-mouth disease (FMD) is considered one of the most contagious diseases of livestock, which can lead to huge economic losses. This disease, present in Africa, Asia and South America, is caused by a virus from the *Picornaviridae* family, genus *Aphthovirus,* referred to as FMD virus (FMDV). Seven different FMDV serotypes have been described (A, O, C, SAT1, SAT2, SAT3 and Asia1). The genome of FMDV comprises a positive-sense single-stranded RNA approximately 8300 nucleotides in length. The viral genome contains a single long ORF, encoding a large polyprotein, further processed into 13 viral mature proteins, whose 4 structural proteins (VP1-VP4). The extensive genetic diversity in FMDV is attributed to the poor proof-reading

*Speaker

†Corresponding author: mathias.vandenbogaert@pasteur.fr

ability of the viral RNA dependent RNA polymerase, with large viral population size and high replication rates. Then, FMDV evolves through genetic drift, where positive selection contributes to fixation of mutations in the capsid coding regions. Although the VP1 coding region of FMDV is useful for isolate characterization, it is relatively short (only _˜8% of the genome length) and, consequently, phylogenetic trees generated from closely related FMDV sequences recovered within outbreak clusters are typically flat, with poor resolution. For this reason, the use of WGS to discriminate between closely related viruses has become commonplace and has subsequently been applied to both human and animal pathogens.

However, incongruences between phylogenies from individual sub-genomic regions suggest that recombination also plays a role in FMDV evolution. Recombination events have indeed been demonstrated within the FMDV genome and have highlighted the fact that particular regions of the FMDV genome appear to be more prone to intertypic recombination than others. The number of exchanges of genome sequences encoding for nonstructural proteins seems to be much more important and numerous, than the events involving the sequences encoding parts of the capsid-coding region. It is therefore important to identify the set of recombination events in FMDV full genome sequences, and to determine the distribution of these events across the FMDV genome. Recombination events are of particular interest as a source for driving FMDV diversity giving rise to FMDV outbreaks. Third-generation sequencing technologies could thus allow to bridge the gap in resolving genome structure uncertainties for such virus.

METHODS

Four isolates of FMDV were sequenced using MiSeq Illumina platform (second-generation) and MinION Oxford Nanopore Technologies (third-generation). Two of these samples were collected from cattle in 2011 from Balochistan Province in Pakistan (PAK-6; PAK-9) and the others originated from Benin (BEN-017, BEN-036) in 2010. The whole genome sequencing (WGS) with Illumina technology were performed using Nextera XT kit in order to produce paired-end reads of approximately 150pb each. The MinION libraries were prepared using 1D2 Sequencing chemistry and Flow cell MIN-10 to obtain one unique long read covering the entire genome of the virus (8Kb).

For second-generation data analysis, the four FMDV genomes were reconstructed using a dedicated pipeline with classic state-of-the-art bioinformatics tools. Third-generation long reads were analyzed using a long reads analysis workflow (including Albacore and Canu Minimap softwares). In both approaches, phylogenetic trees were established using the Mafft tool, allowing to consolidate the geographical origin and the serotype of all isolates and to help solve the recombination events.

A global genomics analysis approach for mapping recombination hotspots appeared to be necessary, particularly for such datasets where the identities of the parental sequences involved in recombination are unknown. More specifically, within the current data study-set, it is generally unknown which FMDV sequence is the recombinant and which is no recombinant.

Mapping of the positions of recombination is done by a phylogenetic-compatibility analysis using phylogeny tree scanning, applied to both publicly available full genomes and newly sequenced isolates. Phylogenetic tree scanning is based on recording the order of each variant in an alignment, giving a successive serie of phylogenetic trees (rooted neighbor-joining trees, 100 bootstrap replicates, and where all branches with $< 70\%$ support are collapsed, moving windows of 300nt and intervals of 100nt), and hence examining the positions in the alignment where phylogenetic relationships change.

To investigate the extent of recombination within the data set, the aligned sequences were examined using the Recombination Detection Program in RDP4, in order to infer breakpoint positions and recombinant sequences for every detected potential recombination event. The results of this analysis are in agreement with the phylogenetic-compatibility analysis in that the distribution of observed breakpoints appears to be non-random.

CONCLUSION.

Incongruent tree topologies between the structural and non-structural coding regions of FMDV isolates suggest that the VP1 phylogeny may not be appropriately reflecting the evolutionary histories of different FMDV isolates. We therefore analyzed the existence of differences in the frequency of recombination between species by an extended comparison of sequences that included all available complete genome sequences available from public databases. Using exhaustive comparisons of fragment sets generated from alignments or the complete genome sequence of the species, it is possible to map regions of phylogenetic incongruity and infer sites of favored recombination using a phylogenetic compatibility matrix (PCM).

The results of these FMDV breakpoint distribution and phylogenetic-compatibility analyses reflect a clear partitioning of structural and non-structural genes in the organization of the genome. This organization facilitates component swapping or recombination that frequently occurs among such viruses.
Confident construction of transmission trees from phylogenetic data, through spatio-temporal epidemiological data, using MinION nanopore sequencing, offers an exciting potential to FMDV diagnostics, and more specifically for resolving recombination scenarios when comparing different field isolates. Such approaches, integrating both novel technological sequencing instruments, together with phylogenetic and epidemiological data, will help understand mechanisms involving the recombination patterns observed in FMDV and other picornaviruses, and will eventually lead to novel insights into epidemiological and phylogeographics issues in FMDV outbreaks.

Keywords: third generation sequencing, oxford nanopore MinION, viral genomic structure, viral recombination

Characterization of DNA replication plasticity among 12 human cell lines

Hadi Kabalane *† 1, Xia Wu 2,3, Malik Kahli 3, Nataliya Petryk 3, Bastien Laperrousaz 1,4, Yan Jaszczyszyn 5, Guenola Drillon 1, Frank-Emmanuel Nicolini 4,6, Aude Robert 7, Cédric Fund 8, Frédéric Chibon 9, Xia Ruohong 2, Joëlle Wiels 7, Françoise Argoul 10, Véronique Maguer-Satta 4, Alain Arneodo 10, Olivier Hyrien 3, Benjamin Audit‡ 1

[1] Univ Lyon, ENS de Lyon, Univ Claude Bernard Lyon 1, CNRS, Laboratoire de Physique, F – 69342, Lyon – France
[2] Physics Department, East China Normal University – Shanghai, China
[3] Institut de Biologie de l'École normale supérieure (IBENS), Département de Biologie, Ecole Normale Supérieure, CNRS, Inserm, PSL Research University, F – 75005 Paris – France
[4] CNRS UMR5286, INSERM U1052, Centre de Recherche en Cancérologie de Lyon, F – 69008 Lyon – France
[5] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Sud, Université Paris – Saclay, Gif-sur-Yvette – France
[6] Centre Léon Bérard, Lyon, F – 69008 – France
[7] UMR 8126, Université Paris-Sud Paris – Saclay, CNRS, Institut Gustave Roussy, Villejuif – France
[8] École normale supérieure, PSL Research University, CNRS, Inserm, IBENS, Plateforme Génomique – 75005 Paris – France
[9] INSERM U1218, Institut Bergonié, F – 3300, Bordeaux – France
[10] LOMA, Université de Bordeaux, CNRS, UMR 5798, Talence, F – 33405 – France

Each cell type presents a specific mean replication timing (MRT) program fingerprint (1). Recently, quantitative analysis of yeast (2) and human (3) genome replication has been achieved by sequencing purified Okazaki fragments (OK-seq), whose strandedness reveal the proportions of rightward- (R) or leftward- (L) moving forks along the genome. Changes in replication fork directionality (RFD = R–L) in turn disclose replication initiation and termination zones as well as regions of unidirectional fork progression. Here, we compared 12 cancer and non-cancer human cell lines using a combination of OK-seq and gene expression (RNA-seq) data (4). This allowed us to question replication plasticity at 10 kb resolution compared the _˜100 kb resolution previously obtained using MRT profiles. Global correlation analysis classified RFD and RNA-seq profiles in accordance to their developmental and/or tumorigenic origins. RFD changes between cell lines are widespread through the genome but more frequent in GC-poor regions. In contrast, RNA-seq changes do not vary uniformly with GC content, indicating that replication changes are dissociated from transcription in a cell-type dependent manner.

Cell lines

The 12 analyzed cell lines include lymphoid, myeloid and adherent cell types. Lymphoid cell

[*] Speaker

[†] Corresponding author: hadi.kabalane@ens-lyon.fr

[‡] Corresponding author: Benjamin.Audit@ens-lyon.fr

lines include two EBV-immortalized lymphoblastoid cell lines (LCLs) (GM06990 and IARC385) and two independently established Burkitt lymphoma cell lines (BLs) (BL79 and Raji). Adherent cells include an epithelial cell line established from a cervix adenocarcinoma (HeLa), two leiomyosarcoma cell lines (LMSs) (TLSE19 and IB118) established from two different patients, and IMR-90 primary human fibroblasts. Myeloid cell lines include a cellular model for establishment and early progression of chronic myeloid leukemia (CML) comprising 2 cell lines and a control (5), and K562, an erythroleukemia cell line derived from a CML patient in blast crisis, which is a late CML model. All analyses were restricted to the 22 autosomes to avoid artefacts due to the XX or XY karyotypes of the studied cell lines.

Cell line classification based on DNA replication and transcription profiling

We used OK-seq to compute RFD profiles in non-overlapping 10 kb windows. To objectively quantify differences between cell lines, we computed the pairwise Pearson correlation coefficients (Cr) between RFD profiles using windows with $>$ 100 OK-seq reads in both cell lines. We ordered them by correlation distance (Dr=1-Cr) using unsupervised hierarchical clustering based on the minimal distance criterion (single linkage clustering). Lymphoid, myeloid and adherent cells formed three separate RFD clusters. Within-group correlation distances are similar so that the three groups are recovered by cutting the classification tree (dendrogram) at level 0.3. Similarly, using RNA-seq data, we estimated the transcriptional level of each gene computing their FPKM (fragments per kilobase of exon model per million mapped fragments) values. Transcriptional correlation (Ct) was computed between log10(FPKM) of genes expressed (FPKM$>$ 1) in both cell lines. Analogous classification was obtained by RNA-seq, except that HeLa clustered with myeloid instead of adherent cells. The correlation coefficients were generally larger by RNA-seq than by RFD. However, the situation is more heterogeneous by RNA-seq where within-group correlation distances increase from lymphoid to myeloid to adherent cells and the three groups cannot be recovered by cutting the dendrogram at a constant level. Within the myeloid group, cell lines clustered in accordance to CML progression using RFD profiles and RNA-seq data, albeit the progression is weaker with the latter. Within the lymphoid cell group, a similar classification of cell lines was also obtained by RNA-seq and RFD. The two BLs (Raji, BL79) were more correlated to each other than to either LCL (GM06990, IARC385), suggesting the existence of BL-specific replication and transcription patterns. Within the adherent cells, different classifications were obtained by RFD and RNA-seq. By RNA-seq, the two LMSs (IB118, TLSE19) were more correlated to each other than to IMR90 and less correlated to HeLa, which in fact clustered with myeloid cells. By RFD, however, the strongest resemblance was observed between TLSE19 and IMR90. The cell of origin and driver mutations of LMSs are currently unclear. These results may help to distinguish different types of LMS and suggest a possible differentiation of TLSE19 and IB118, which were derived from a buttock muscle tumor and a scalp tumor, respectively. A possible interpretation is that the strong correlation of the RNA-seq profiles of the two LMSs reflects the selection for a cancerous phenotype, whereas the RFD patterns more predominantly reflect their different cell type of origin.

In summary, the global correlation analysis clustered the RFD profiles of the 12 cell lines in accordance to their developmental origin and/or cancerous character, reflecting progression along specific tumour progression pathways. Globally similar results were obtained by RNA-seq, but divergences between RNA-seq and RFD classifications were also observed. These results suggest that recurrent replication changes occur in specific tumour types but that the tightness of their connection with transcription changes may depend on the cellular context.

Cell line classification does not result from localized changes

We investigated whether the differences among RFD profile of the 12 cell lines were caused by changes in specific regions or if they are widespread along the genome. First, we repeated the previous analyses for each chromosome separately and observed that the classification obtained for the entire genome was recapitulated for each separate chromosome, with minor exceptions. Second, since GC-content fluctuations recapitulate the non uniform organization of gene size (6), gene density (7), gene expression (8) and replication timing (9) along the human genome, we repeated the analyses separately for five increasing GC-content classes following the 5 isochores classification of the human genome (10) in light isochores L1 (GC$<$ 37%) and L2 (37%$\leq$GC$<$ 41%) and heavy isochores H1 (41%$\leq$GC$<$ 46%), H2 (46%$\leq$GC$<$ 53%) and H3 (GC$\geq$53%). A similar hierarchical clustering of cell lines to that obtained with genome-wide correlation analysis was recovered in each GC-content class, suggesting that RFD changes between cell lines are widespread through the five isochores. Third, to assess this robustness at a higher resolution, we adopted a bootstrap approach on RFD profiles. We generated a large number of random probes (1000 per probe size), 50 kb to 50 Mb in size, consisting each of 5 to 5,000 randomly selected 10 kb windows. For each probe, we computed all pairwise RFD correlation coefficients and their correlation (Cp) with the global genome correlation values. We observed that a random probe $\geq$5 Mb allows (i) the faithful reconstruction of the global correlation values (Cp$>$ 0.93) and (ii) the correct classification of the cell lines in the lymphoid, myeloid and adherent groups with a probability $>$ 0.95. This demonstrates that cell-line specific RFD changes are widely distributed over the entire genome. Cell line classification does not result from localized regions but is representative of the global genome.

Replication changes are stronger in GC-poor regions independently of transcriptional changes

Although identical cell line classification was obtained for each GC-content class, we observed a coupling between RFD changes and GC content. Pairwise RFD correlation coefficients increased with GC content most of the time. When the pairwise correlation coefficient differences between each GC-content class and the entire genome were computed, most differences were negative in L1, null in L2 and increasingly positive in H1 to H3. In other words, the RFD profiles were less, equally, or more similar to each other in the L1, L2, or H1-3 fractions, respectively, than in the global genome. Therefore, RFD changes are more frequent in the GC-poor fractions of the genome. These observations were not due to a higher technical noise in GC-poor regions. If due to noise differences, correlations differences should vanish when the scale of analysis is increased. However, the cell classification and the GC dependence of correlation differences were conserved or even enhanced when the scale of analysis was increased from 10 kb to 100 kb, 200 kb and 1Mb. A similar GC-content analysis was performed with the RNA-seq data. GC content-dependent changes in correlation coefficients were less marked than for RFD. Unlike RFD, correlation difference matrices of RNA-seq data showed no general tendency to follow GC content. For example, inside a given group (lymphoid or myeloid or adherent), the tendency was similar to RFD, but an opposite tendency was observed for the lymphoid vs. myeloid comparisons and the lymphoid vs. adherent comparisons did not reveal a group tendency. A number of comparisons were maximum for intermediary GC content. This suggests that replication changes are at least partly dissociated from transcription changes, to an extent that depends on the cellular context.

Regions of stable RFD are replicated early and highly expressed

To identify regions where the RFD profiles are particularly stable or variable, we computed the difference between the pairwise RFD correlation coefficients in each non-overlapping 5 Mb windows and the global pairwise correlation values. Then, for each window, we derived a Z-score as the ratio between the mean and standard deviation of the pairwise correlation differences. Stable (resp. variable) RFD regions with mostly positive (negative) pairwise correlation differences were selected as the windows with the 5% highest (resp. lowest) Z-score values (z> 1.64 and z< -1.25, resp.). Observation of the 12 RFD profiles in the selected regions confirmed the effectiveness of the methodology. The regions with stable (resp. variable) RFD profiles are associated to GC-rich (resp. AT- rich) regions (median GC content is 0.42 (resp. 0.36)) as expected from the GC class analysis. In the same manner, the stable (resp. variable) regions are associated with high (resp. low) level of expression. For example, considering RNA-seq data in Raji, 50% of the 200 kb windows within stable regions have an average FPKM> 1 whereas this proportion drops to 10% for variable regions. Finally, using available MRT data in GM06990, we observed that the median MRT of 200 kb windows in the variable regions is 0.7 significantly higher that the value 0.4 for stable regions. These results underline that regions of stable RFD tend to be in early replicating and highly transcribed regions. It provides further evidence that RFD changes are to some extent disconnected from the transcriptional program.

Conclusion

A global, unbiased correlation approach revealed that the RFD profiles of 12 cancer and non-cancer cell types cluster in three separate groups corresponding to lymphoid, myeloid and adherent cells. Therefore, cancer-associated changes in replication do not blur their developmental origin signature. The global correlation analyses further revealed that RFD changes between cell lines are widespread through the genome but more frequent in GC-poor regions. In contrast, RNA-seq changes do not vary uniformly with GC content. Changes in replication program predominantly target GC-poor, lowly expressed and late replicating regions. These results strengthen the notion that replication changes are dissociated from transcription changes, to an extent that specifically depends on the compared cell types.

Acknowledgments

HK acknowledges support from the Agence Nationale de la Recherche (ANR-15-CE12-0011-01) and the Fondation pour la Recherche Médicale (FRM DEI201512344404).

References

1. Ryba, T. et al. (2011) PLoS Comput. Biol. 7, e1002225.

2. McGuffee, S.R et al. (2013) *Mol Cell*, **50**, 123-135.

3. Petryk, N et al. (2016) *Nature communications*, **7**, 10208.

4. Wu, X. et al. (2018) *Nucleic Acids Res.* in revision.

5. Laperrousaz, B et al. (2013) *Blood*, **122**, 3767-3777.

6. Duret, L. et al. (1995) *J Mol Evol*, **40**, 308-317.

7. Lander, E.S. et al. (2001) *Nature*, **409**, 860-921.

8. Woodfine, K. et al. (2005) *Cell Cycle*, **4**, 172-176.

9. Woodfine, K. et al. (2004) *Hum Mol Genet*, **13**, 191-202.
10. Bernardi, G. (2001) *Gene*, **276**, 3-13.

Keywords: DNA replication program plasticity, Replication Fork Directionality, Okazaki Fragments, GC content, Mean Replication Timing, RNA seq.

Structural Bioinformatics
Proteomics

Assessing the functional impact of genomic alterations using proteogenomics

Georges Bedran *† 1,2,3, Yves Vandenbrouck 4, Lucid Belmudes 4, Eric Bonnet‡ 5, Jean-François Deleuze 5, Delphine Pflieger 4, Christophe Battail 4

[1] CEA/DRF/BIG/BGE/EDYP (Exploring of the Dynamics of Proteomes) – Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble – 17 Avenue des Martyrs, F-38054 Grenoble, France

[2] UFR Sciences et Techniques - Université de Rouen (Master de bioinformatique) – UFR Sciences et Techniques - Université de Rouen – Place Emile Blondel, 76821, Mont-Saint-Aignan, France, France

[3] CEA/DRF/JACOB/CNRGH (Centre National de Recherche en Génomique Humaine) – CEA Evry 2 rue Gaston Crémieux 91006 Evry cedex – France

[4] CEA/DRF/BIG/BGE/EDYP (Exploring of the Dynamics of Proteomes) – Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble – 17 rue des Martyrs, 38054, Grenoble, France, France

[5] CEA/DRF/JACOB/CNRGH (Centre National de Recherche en Génomique Humaine) – CEA Evry 2 rue Gaston Crémieux 91006 Evry cedex – 2 rue Gaston Crémieux, 91057, Evry, France, France

In recent years, high-throughput DNA sequencing technologies have allowed consortia such as the Cancer Genome Atlas Research Network (TCGA) to map genetic alterations from hundreds of tumor biopsies of different cancers [1]. These studies established the complex landscape of genetic alterations found in cancers, notably made of single-nucleotide variants (SNVs), insertions and deletions (indels), aberrant gene fusions and alternative splice variants. These profiles were explored to work on the identification of mutations responsible for tumorigenesis, to classify tumors for better diagnostics and to identify potential therapeutic targets. However, proteins are the main players of the cellular function, and aberrant proteins drive tumor initiation, progression and response to treatment. In addition, beyond the information provided by the genome and transcriptome knowledge, the CPTAC (Clinical Proteomic Tumor Analysis Consortium) has recently demonstrated on tumor biopsies that the integration of proteome profiling, obtained by mass spectrometry, revealed perturbations inaccessible to genomics alone and to reveal new tumor subtypes [2].
A major question still poorly addressed in cancer biology is how the information flow from genome to transcriptome to proteome. Given the low correlation level between mRNA and protein abundances, the integrated genomic, transcriptomic and proteomic views of the same biological samples is the best strategy to explore this question [3].

Recent technological and methodological advances in tandem mass spectrometry coupled with liquid chromatography (LC-MS/MS) have greatly improved coverage of complex protein samples, and increased measurement precision. In proteomics, peptides are most commonly identified using a shotgun approach by matching MS/MS spectra against theoretical spectra of all

*Speaker

†Corresponding author: georges.bedran@etu.univ-rouen.fr

‡Corresponding author: eric.bonnet@cng.fr

candidate peptides represented in a generalist protein sequence reference database[4]. It is at this stage of database searching that proteomic approaches lose the capability to identify genetic aberrations carried by proteins. Thus, the first step toward 'personalized' proteomics for cancer studies, where each sample has its own universe of mutations, is leveraging other omic data in approaches called proteo-genomics, and to use them as a priori information to create customized protein sequence databases [5].

The customized protein database can be generated using multiple data sources such as six-frame translation of the genome, ab initio gene prediction, annotated genes or from expressed genes and their genetic alterations. The optimal choice depends on the goals of the experiments; more specifically, on the types of novel peptides that the study seeks to identify. However, searching MS/MS spectra against large protein databases may result in a lack of sensitivity and specificity in peptide identification. Thus, a key consideration in proteo-genomics is the selection of the optimal strategy for generating the customized sequence database, i.e., finding the right balance between the completeness of the database and its size [6].

Cancer cells harbor a massive amount of punctual mutations as well as deletions and insertions. Thus, the use of RNA sequencing data to construct a personalized protein database, from expressed normal and mutated transcripts, allows the detection of peptides covering novel SAAVs (single amino acid variants), INDELs, novel exon-exon splice junctions and fusion genes while keeping a small database size.

Despite the technological advances in LC-MS/MS and the use of an optimal personalized search space (database) far from all mutated peptides in a sample are detected by discovery proteomics. There are two main reasons for this: the large dynamic range in protein abundance and the lack of selection of all parent ions for fragmentation. To fix this problem, a combination of shotgun and targeted proteomics can be used to increase the number of mutated peptides accessible to proteomic characterization. In the targeted proteomic approach, the LC-MS/MS system will focus its analysis on a list of peptides of interest. This will improve greatly the capability to detect and quantify them in comparison to discovery proteomics.

We designed and implemented a bioinformatic methodology in proteo-genomics dedicated to the detection and the quantification of cancer genetic variants by discovery and targeted proteomics.

The methodology starts from the list of genetic variants (Single Amino-Acid Variants or SAAVs, insertions/deletions, and novel splicing junctions) and transcripts abundance values obtained from the RNA-seq profiling of a biological sample. These information are used to generate a customized protein database using The R package customProDB [5] serving both the discovery and the targeted proteo-genomic approaches.

In discovery proteo-genomics, the experimental spectra are matched to tryptic peptides that derive from the generated protein database using a proteomic search engine (MS-GF+) [7]. Peptides are then validated using a target-decoy strategy [8] to control the false discovery rate (FDR) (i.e controlling the FDR threshold to 1%) and separated into 4 categories: Normal, saav, indels, novel exon-exon splicing junctions. Peptides carrying SAAVs were further studied to determine if the observed mass shift between the mutated peptide and the wild type amino acid does not match the mass of one of the common chemical or post-translational modifications (e.g. oxidation, deamidation, carbamylation, acetylation, etc.). This method uses a comprehensive database of protein modifications for mass spectrometry called unimod [9].

The selection of the shortest list of proteins groups that can explain all of the identified peptides of a discovery analysis, referred to as protein inference, is a critical step in proteomic studies. This approach suffers from the difficulty of correctly assigning peptides, shared with several proteins, to the correct protein group. We decided to integrate a gene inference strategy which consists of mapping each peptide to the human genome sequence in order to obtain its corresponding genomic coordinates and then estimate expressed genes list using the R package proBAMtools [10]. Moreover, this proBAM (protein bam) format allows the integrated exploration of transcriptomic and proteomic information using a genome browser such as the integrative genomics viewer (IGV).

The targeted proteo-genomic approach consists of generating features related to each mutated variant to assess their detectability in proteomics. These peptide features consist of the protein tryptic digestion, the length, the ionization, the expression of the corresponding transcript obtained by RNA-seq and the mapping to a unique gene. Beyond the proteomic detectability prediction, the methodology goes further to also annotate each genetic variant according with other genomic databases such as dbSNP, 1000G for polymorphisms, Cosmic DB for cancer related status and dbNSFP for predicting the deleterious impact on the protein using oncotator [11]. These properties are leveraged not only to target and identify these peptides but also to quantify their abundance.

We have so far applied this bioinformatic methodology using transcriptomic and discovery proteomic data produced by the mRNA sequencing and the LC-MS/MS profiling of HCT-116 cell line (human colorectal cancer).

The RNA sequencing resulted in 2 x 40M reads mapped to the human genome (hg19).The genetic variant calling allowed us to notably detect 3820 SAAVs with associated transcriptional expression greater than 0.2 RPKM.

These variants were used to generate a customized protein database consisting of 59902 wild type isoform protein sequences, 11031 protein sequences with SAAVs, 820 protein sequences with indels and 5005 peptides covering novel exon-exon splicing junctions. Based on this search space, we interrogated the discovery proteomics data of HCT-116 and identified 27280 peptide spectrum matches, including 127 SAAVs, of which 40 were also found by CPTAC colorectal cancer study [2], 1 indel along with 13 novel exon-exon splice junction peptides. Our gene inference procedure identified 4761 expressed gene groups of which 132 contained more than 1 gene. We found using our discovery proteo-genomic approach that only 4% of SAAVs identified by RNA-seq profiling of HCT116 are detected by discovery proteomics. However, the targeted proteo-genomic approach shows that two-third of SAAVs identified at transcriptome level are theoretically detectable in proteomics.

We are currently investigating these genetic variants in order to:

(i) to confirm the identification of SAAVs detected from discovery proteomics
(ii) to explore a connected network of kinases and phosphatases mutated in HCT116
(iii) to explore the allelic-expression of these SAAVs and assess if the allelic ratio is conserved or perturbed between the transcriptome and proteome levels.

To conclude, our bioinformatic methodology will be accessible to the research community under an open source license. It is implemented using mostly Python and has some R dependencies. Furthermore, it is multithreaded and multiprocessed to allow the proteo-genomic analysis of large cohorts of tumoral biopsies using HPC infrastructure. We will evaluate its scalability using cohorts of tumoral biopsies profiled in proteo-genomics by TCGA/CPTAC consortia.

Keywords

Proteogenomics ; RNA-sequencing ; colorectal cancer, SAAVs (single amino acid variants) ; splice junctions ; INDELs (insertions / deletions) ; discovery proteomics ; targeted proteomics ; open source ; bioinformatics.

References

1. The Cancer Genome Atlas Network. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. Nature. 2012;487:330–7.

2. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014;513:382–7.

3. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell. 2016;165:535–50.

4. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. Journal of proteomics. 2010;73:2092–123.

5. Wang X, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics [Internet]. 2013; Available from: http://bioinformatics.oxfordjournals.org/content/early/2013/09/20/bioinformatics.btt543.abstract

6. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Meth. 2014;11:1114–25.

7. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun [Internet]. 2014;5. Available from: http://dx.doi.org/10.1038/ncomms6277

8. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Meth. 2007;4:207–14.

9. Creasy DM, Cottrell JS. Unimod: Protein modifications for mass spectrometry. PROTEOMICS. 2004;4:1534–1536.

10. Wang X, Slebos RJC, Chambers MC, Tabb DL, Liebler DC, Zhang B. proBAMsuite, a Bioinformatics Framework for Genome-Based Representation and Analysis of Proteomics Data. Molecular & Cellular Proteomics: MCP. 2016;15:1164–75.
11. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, et al. Oncotator: cancer variant annotation tool. Human mutation. 2015;36:E2423-9.

Keywords: Proteogenomics, RNA, sequencing, colorectal cancer, SAAVs (single amino acid variants), splice junctions, INDELs (insertions / deletions), discovery proteomics, targeted proteomics, open source, bioinformatics

Structural and dynamics studies of a potassium channel and disease-associated mutants

Charline Fagnen *† [1,2], Iman Oubella [2], Yasmina Mhoumadi [1,2], Aline De Araujo [1,2], Eric Forest [3], David Perahia [4], Catherine Vénien-Bryan‡ [5]

[1] Laboratoire de Biologie et de Pharmacologie Appliquée (LBPA) – École normale supérieure - Cachan, Centre National de la Recherche Scientifique : UMR8113 – 61 AVENUE DU PRESIDENT WILSON 94235 CACHAN CEDEX, France

[2] Institut de minéralogie, de physique des matériaux et de cosmochimie (IMPMC) – Institut de recherche pour le développement [IRD] : UR206, Université Pierre et Marie Curie (UPMC) - Paris VI, CNRS : UMR7590, Muséum National d'Histoire Naturelle (MNHN) – Tour 23 - Barre 22-23 - 4e étage - BC 115 4 place Jussieu 75252 PARIS, France

[3] Institut de Biologie Structurale – CEA, CNRS : UMR5075, UJF – Grenoble, France

[4] Laboratoire de Biotechnologie et Pharmacologie génétique Appliquée (LBPA) – CNRS : UMR8113, École normale supérieure (ENS) - Cachan – 61 AVENUE DU PRESIDENT WILSON 94235 CACHAN CEDEX, France

[5] Institut de minéralogie, de physique des matériaux et de cosmochimie (IMPMC) – Museum National d'Histoire Naturelle, Université Pierre et Marie Curie - Paris 6 : UM120, Institut de recherche pour le développement [IRD] : UR206, Centre National de la Recherche Scientifique : UMR7590 – Tour 23 - Barre 22-23 - 4e étage - BC 115 4 place Jussieu 75252 PARIS, France

Inwardly-rectifying potassium (Kir) channels are transmembrane proteins that play a key-role in many physiological processes such as the creation and the propagation of the neuronal action's potential, the regulation of cellular volume, the muscular contraction and the cardiac pulse. Their physiological importance is highlighted by the fact that genetically inherited defects in Kir channels are responsible for a wide-range of channelopathies1 including Andersen's syndrome2–4. To date unfortunately, this disease does not have any effective treatment. To elucidate how channel function becomes defective in the disease state requires a detailed understanding of how the channel goes from the open to the closed states. This will allow the identification of the most suitable regions and motions for the binding of small correctors or drug which could influence the conformation of intracellular gating elements.
In this work we are focusing on the Kir2.1 channel and three important mutations (G144S, C154Y, R312H) which have been responsible for dysfunction of the channel (loss of function) leading to the rare disease Andersen syndrome. The atomic structures of the open states KirBac3.1 (bacterial homologue of the human Kir2.1) and several of its mutants were solved by our team5. These structures suggested that a rotation of the cytoplasmic domain could be associated with the opening of the channel. In order to test this "twist to open hypothesis"5, MDeNM6 (Molecular Dynamic using Excited Normal Modes) simulations were performed. The MDeNM method consists in combining two simulations methods: molecular dynamics and nor-

*Speaker

†Corresponding author: charline.fagnen@upmc.fr

‡Corresponding author: catherine.venien-bryan@upmc.fr

mal modes. The former allows us to observe particularly fast and small amplitude movements such as side-chain movements or loops, the latter on the other hand describes slow and collective movements of large amplitude. This mixed approach gives access to a wider exploration of the conformational space and allows moreover to determine the conformational populations of the different states (open and closed).

Interestingly, during these *in silico* MDeNM studies focusing on the rotation of the cytoplasmic domain, the simulation revealed a slight opening of the channel but not sufficiently pronounced to support the idea that this rotation has a central role in the opening and closing or gating mechanism of the channel. Following this observations, we investigated the modes describing the opening of the channel and made MDeNM structures from them in order to: i) obtain start structures for classical molecular dynamics to explore the conformational space, ii) identify the most involved parts of the protein in the transition from the open to closed state.

To check our theoretical results, we used experimental data such as structural analysis using cryo-electron microscopy or mass spectrometry by H/D exchange (HDX-MS)7. Preliminary results show a kink8 on the helix of the channel in agreement which is visible in electron crystallography and X-ray crystallography (5). HDX-MS is an experimental tool which has been used here to determinate the local flexibilities of the protein. The method consists in replacing proton of the protein by deuterium, these exchange are timed. The longer the exchanges take place, the more rigid the studied region is.. We found a good correlation between these two methods, indeed, the most flexible regions in HDX-MS correspond with the region with the more fluctuations in MDeNM. These preliminary data allow us to validate the use of MDeNM and the choice of the modes to describe the opening of the channel.

We pursue and launch MDeNM more accurate simulations. We hope to identify residues or regions involve in the molecular mechanism of gating and understand how this gating works. In parallel, as the human Kir2.1 structure is not resolved yet, we have started image analysis on electron cryomicroscopy data and obtained a first map about 7Å. New data collected by Titan Krios (300keV) microscope are under analysis in order to access to high-resolution structure of Kir2.1.

References

(1) Abraham, M. R.; Jahangir, A.; Alekseev, A. E.; Terzic, A. Channelopathies of Inwardly Rectifying Potassium Channels. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **1999**, *13* (14), 1901–1910.

(2) Plaster, N. M.; Tawil, R.; Tristani-Firouzi, M.; Canún, S.; Bendahhou, S.; Tsunoda, A.; Donaldson, M. R.; Iannaccone, S. T.; Brunt, E.; Barohn, R.; Clark, J.; Deymeer, F.; George, A. L.; Fish, F. A.; Hahn, A.; Nitu, A.; Ozdemir, C.; Serdaroglu, P.; Subramony, S. H.; Wolfe, G.; Fu, Y. H.; Ptácek, L. J. Mutations in Kir2.1 Cause the Developmental and Episodic Electrical Phenotypes of Andersen's Syndrome. *Cell* **2001**, *105* (4), 511–519.

(3) Tristani-Firouzi, M.; Jensen, J. L.; Donaldson, M. R.; Sansone, V.; Meola, G.; Hahn, A.; Bendahhou, S.; Kwiecinski, H.; Fidzianska, A.; Plaster, N.; Fu, Y.-H.; Ptacek, L. J.; Tawil, R. Functional and Clinical Characterization of KCNJ2 Mutations Associated with LQT7 (Andersen Syndrome). *J. Clin. Invest.* **2002**, *110* (3), 381–388.

(4) Hosaka, Y.; Hanawa, H.; Washizuka, T.; Chinushi, M.; Yamashita, F.; Yoshida, T.; Komura, S.; Watanabe, H.; Aizawa, Y. Function, Subcellular Localization and Assembly of a Novel Mutation of KCNJ2 in Andersen's Syndrome. *J. Mol. Cell. Cardiol.* **2003**, *35* (4),

409–415.

(5) Bavro, V. N.; De Zorzi, R.; Schmidt, M. R.; Muniz, J. R. C.; Zubcevic, L.; Sansom, M. S. P.; Vénien-Bryan, C.; Tucker, S. J. Structure of a KirBac Potassium Channel with an Open Bundle Crossing Indicates a Mechanism of Channel Gating. *Nat. Struct. Mol. Biol.* **2012**, *19* (2), 158–163.

(6) Costa, M. G. S.; Batista, P. R.; Bisch, P. M.; Perahia, D. Exploring Free Energy Landscapes of Large Conformational Changes: Molecular Dynamics with Excited Normal Modes. *J. Chem. Theory Comput.* **2015**, *11* (6), 2755–2767.

(7) Mehmood, S.; Domene, C.; Forest, E.; Jault, J.-M. Dynamics of a Bacterial Multidrug ABC Transporter in the Inward- and Outward-Facing Conformations. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (27), 10832–10836.

(8) Kuo, A.; Domene, C.; Johnson, L. N.; Doyle, D. A. ; Vénien-Bryan, C. Two different conformational states of the KirBac3.1 potassium channel revealed by electron crystallography. *Struct. Lond. Engl. 1993* **2005**, 13, 1463–1472.

Keywords: MDeNM, transmembrane protein, Kir, channel, normal modes, molecular dynamics

Protein interaction energy landscapes are shaped by functional and also non-functional partners

Hugo Schweke *† [1], Marie-Hélène Mucchielli-Giorgi [1], Sophie Sacquin-Mora [2], Wanying Bei [1], Anne Lopes‡ [1]

[1] Institute for Integrative Biology of the Cell (I2BC) – Université Paris Sud - Paris XI, CEA, Centre national de la recherche scientifique - CNRS (France) – France
[2] Laboratoire de biochimie théorique (LBT) – CNRS : UPR9080 – 13 Rue Pierre et Marie Curie 75005 PARIS, France

Biomolecular interactions are central for many physiological processes and are of utmost importance for the functioning of the cell. Particularly protein-protein interactions have attracted a wealth of studies these last decades (Janin et al., 2008; Robinson et al., 2007). The concentration of proteins in a cell has been estimated to be approximately 2-4 million proteins per cubic micron (Milo, 2013). In such a highly crowded environment, proteins constantly encounter each other and numerous non-specific interactions are likely to occur (McGuffee and Elcock, 2010). For example, in the cytosol of *S. cerevisiae* a protein can encounter no less than 2000 different proteins (Levy et al., 2014). In this complex jigsaw puzzle, each protein has evolved to bind the right piece in the right way (positive design) and to prevent misassembly and non-functional interactions (negative design) (Garcia-Seisdedos et al., 2017; Pechmann et al., 2009).
Consequently, positive design constrains the physico-chemical properties and the evolution of protein-protein interfaces. Indeed, a strong selection pressure operates on binding sites to maintain the functional assembly. For example, homologs sharing at least 30% sequence identity almost invariably interact in the same way (Aloy et al., 2003). Conversely, negative design prevents proteins to be trapped in the numerous competing non-functional interactions inherent to the crowded environment of the cell. Particularly, the misinteraction avoidance shapes the evolution and physico-chemical properties of abundant proteins, resulting in slower evolution and less sticky surfaces than what is observed for less abundant ones (Levy et al., 2012; Yang et al., 2012). The whole surface of abundant proteins is thus constrained, preventing them to engage deleterious non-specific interactions that could be of dramatic impact for the cell at high concentration (Levy et al., 2012). Recently, it has been shown in *E. coli* that the net charge as well as the charge distribution on protein surfaces affect the diffusion coefficients of proteins in the cytoplasm (Schavemaker et al., 2017). Positively charged proteins move up to 100 times more slowly as they get caught in non-specific interactions with ribosomes which are negatively charged and therefore, shape the composition of the cytoplasmic proteome (Schavemaker et al., 2017).

All these studies show that both positive and negative design effectively operate on the whole

*Speaker

†Corresponding author: hugo.schweke@i2bc.paris-saclay.fr

‡Corresponding author: anne.lopes@i2bc.paris-saclay.fr

protein surface. Binding sites are constrained to maintain functional assemblies (i.e. functional binding modes and functional partners) while the rest of the surface is constrained to avoid non-functional assemblies. Consequently, these constraints should shape the energy landscapes of functional and non-functional interactions so that non-functional interactions do not prevail over functional ones. This should have consequences (i) on the evolution of the propensity of a protein to interact with its environment (including functional and non-functional partners) and (ii) on the evolution of the interaction propensity of the whole surface of proteins, non-interacting surfaces being in constant competition with functional binding sites. Concretely, we can hypothesize that the interaction propensity of the whole surface of proteins is constrained during evolution in order to (i) ensure that proteins correctly bind functional partners, and (ii) limit non-functional assemblies as well as interactions with non-functional partners.

In this work, we focus on protein surfaces as a proxy for functional and non-functional protein-protein interactions. We interrogate how this competition constrains the behavior of proteins with respect to their partners or random encounters with a novel theoretical framework based on an original representation of interaction energy landscapes. These latters are represented with two-dimensional (2D) energy maps that reflect in a synthetic way the propensity of a protein to interact. Docking algorithms are now fast enough for large-scale applications and allow for the characterization of interaction energy landscapes for thousand of protein couples. In particular, docking simulations enable to energetically characterize all possible interactions involving functional but also arbitrary partners, and thus to simulate the interaction of arbitrary partners which is very difficult to address with experimental approaches. Recently, we and others have demonstrated through extensive cross-docking experiments that the docking of functional but also arbitrary protein pairs is a viable route to predict protein binding sites as well as protein partners (Wass *et al*, 2011, Lopes *et al*, 2013, Ohue *et al*, 2013, Vamparys *et al*, 2016). Here, we take advantage of these recent advances and move forward by studying the evolution of interaction energy landscapes involving either true partners or arbitrary protein pairs. Therefore, we performed several thousands of cross-docking simulations to systematically compare the resulting interaction energy maps of a given protein docked with different sets of homologs, corresponding to its functional partner's family or arbitrary protein families. To quantify the conservation during evolution of the whole surface's propensity of the protein to interact with a protein family, we computed an AUC (Area Under the Roc Curve) estimating our capacity to retrieve protein families based only on the energy maps. The AUC is very high, of 80%, showing that the interaction propensity of the whole protein surface is conserved for homologous partners, be they functional or not. Strikingly, the predictive power of particular protein surface regions, which we define based on interaction energy levels, is as high as that of the whole surface.

We reveal protein surface properties that allow for the proposition of a new model of protein surfaces. We demonstrate that our 2D energy maps based strategy makes possible in an efficient and automated way, to extract from the whole surface of proteins, information relevant to protein interactions. While most studies aiming at depicting protein-protein interactions focus on native binding sites of proteins, we bring a new perspective on protein-protein interactions with the physical characterization of not only known binding sites, but also of the rest of the protein surface. The latter is known to play an important role in protein interactions by constantly competing with the formers. We show that the interaction propensity of the rest of the protein surface is not homogenous and that the whole protein surface comprises regions of different binding energy levels (i.e. hot, intermediate and cold regions for favorable, intermediate and unfavorable interaction regions respectively) (i) whose localizations are specific to the protein partner family (ii) and which display specific structural and physico-chemical properties. We propose a new model of protein surfaces where protein surface regions, in the crowded cellular environment, serve as a proxy for regulating the competition between functional and non-functional interactions. In this model, intermediate and cold regions play an important and so far undermined role by preventing non-functional interactions (i.e. non-functional binding modes and non-functional protein pairs) and thus guiding the interaction process toward functional interactions. Hot regions can then select the functional interaction (i.e. functional assembly with the functional partner) among the competing ones through interfaces optimized for the native partner.

Our theoretical framework opens the way to a variety of applications related to protein structure and function. We show that our framework enables to highlight and characterize hot regions on a protein surface, which can be either specific or conserved for all partners, and allows for the development of novel methods for protein binding sites prediction and classification as functional or promiscuous. Our 2D energy maps based framework provides an entry point for further protein functional characterization as it reveals biophysical and functional protein properties that could not have been revealed with classical descriptors such as RMSD or sequence identity. One should notice that the dataset set up in this work had to fulfill protein structure and sequence homology constraints and thus does not represent a real crowded cellular environment. Nevertheless, we show that our strategy enables to explore the propensity of a protein to interact with hundreds of selected partners, thus addressing the behavior of a protein in a specific cellular environment. It goes beyond the classical use of binary docking to provide a systemic point of view of protein interactions with a residue resolution, and thus opens the way to further developments for the characterization and understanding of protein function in a crowded environment.

References

Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003;332:989–98.

Garcia-Seisdedos H, Empereur-Mot C, Elad N, Levy ED. Proteins evolve on the edge of supramolecular self-assembly. Nature 2017;548:244–7. doi:10.1038/nature23320.

Janin J, Bahadur RP, Chakrabarti P. Protein-protein interaction and quaternary structure. Q Rev Biophys 2008;41:133–80. doi:10.1017/S0033583508004708.

Levy ED, De S, Teichmann SA. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. Proc Natl Acad Sci USA 2012;109:20461–6. doi:10.1073/pnas.1209312109.

Lopes A, Sacquin-Mora S, Dimitrova V, Laine E, Ponty Y, Carbone A (2013) Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. PLoS Comput Biol 9(12):e1003369.

McGuffee SR, Elcock AH. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. PLoS Comput Biol 2010;6:e1000694. doi:10.1371/journal.pcbi.1000694.

Milo R. What is the total number of protein molecules per cell volume? A call to rethink some published values. Bioessays 2013;35:1050–5. doi:10.1002/bies.201300066.

Ohue M, Matsuzaki Y, Shimoda T, Ishida T, Akiyama Y (2013) Highly precise protein-protein interaction prediction based on consensus between template-based and de novo docking methods. BMC Proceedings (BioMed Central), p S6.

Pechmann S, Levy ED, Tartaglia GG, Vendruscolo M. Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. Proc Natl

Acad Sci USA 2009;106:10159–64. doi:10.1073/pnas.0812414106.

Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. Nature 2007;450:973–82. doi:10.1038/nature06523.

Schavemaker PE, Śmigiel WM, Poolman B. Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome. Elife 2017;6. doi:10.7554/eLife.30084.

Vamparys L, Laurent B, Carbone A, Sacquin-Mora S (2016) Great interactions: How binding incorrect partners can teach us about protein recognition and function. Proteins Struct Funct Bioinforma 84(10):1408–1421.

Wass MN, Fuentes G, Pons C, Pazos F, Valencia A (2011) Towards the prediction of protein interaction partners using physical docking. Mol Syst Biol 7(1):469.
Yang J-R, Liao B-Y, Zhuang S-M, Zhang J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci USA 2012;109:E831-840. doi:10.1073/pnas.1117408109.

Keywords: protein protein interactions, interaction propensity, interaction energy landscapes, non, functional interactions, protein surfaces

NR-DBIND : A database dedicated to nuclear receptor binding data including negative data and pharmacological profile

Manon Réau * [1], Matthieu Montes† [2]

[1] Laboratoire Génomique, Bioinformatique et Applications (GBA) – Conservatoire National des Arts et Métiers - CNAM (FRANCE) – 2 rue Conté, 75003 Paris, France

[2] Laboratoire Génomique, Bioinformatique et Applications (GBA) – Conservatoire National des Arts et Métiers - CNAM (FRANCE) – 2 rue Conté, 75003 Paris, France

Nuclear receptors (NRs) are transcription factors capable of regulating gene expression in various key physiological processes through their interaction with small hydrophobic molecules. They constitute an important class of targets for drugs and endocrine disruptors and they are widely studied for both human health and environmental risks. A major focus in the study of NRs is to identify selective modulators with reduced side effects and to evaluate NR-related chemicals endocrine-disrupting potential. Today, the NR family is among the most studied protein families, and the quantity of experimental binding and activity data published in the literature should be valuably used to boost NRs compounds profiling, ligand-based and structure-based drug design, and SAR studies. In the present work, we gathered diverse NR experimental data that has been published in the literature in a single database named Nuclear Receptor DataBase Including Negative Data (NR-DBIND) to help extracting qualitative information for chemists, biologists and toxiciologists. All data has been manually curated through literature proof reading and particular effort was invested on defining homogeneous and unbiased subsets. NRs were included in the database if at least one non-mutated and documented structure was available in the Protein Data Bank. A NR ligand was included in its corresponding NR dataset if its affinity for the corresponding NR was documented in the literature. When available, activity data was added for pharmacological profile assignment. In total, 15116 interaction data were collected for 28 NRs, corresponding to 13566 unique ligand/protein pairs, including literature reported negative data. The NR-DBIND is freely available at http://www.nr-dbind.drug-design.fr and proposes multiple datasets. pIC50 and pKi affinity values are considered separately, and 3 subsets are provided depending on the level of accuracy of the pharmacological profiling annotation. To date, the NR-DBIND constitutes the largest annotated database on NRs ligands and structures, represents a robust basis 1. for the calibration and benchmark of Computer Aided Drug Design methods; 2. for the identification of new NR modulators and 3. for the assessment of environmental risks linked to endocrine disruptors.

Keywords: Nuclear Receptor, binding data, pharmacological profile, Computer Aided Drug Design,

*Speaker

†Corresponding author: matthieu.montes@cnam.fr

benchmark, Structure Activity Relationship, endocrine disruptors

A hybrid combinatorial method for docking a single-stranded RNA in a protein pocket at the thermodynamic equilibrium

Chinmay Singhal [1], Yann Ponty [2,3], Isaure Chauvot De Beauchêne * [4]

[1] AMIBio team, INRIA Saclay – INRIA – France
[2] Laboratoire d'informatique de l'école polytechnique [Palaiseau] (LIX) – CNRS : UMR7161, Polytechnique - X – Route de Saclay 91128 PALAISEAU CEDEX, France
[3] AMIB (INRIA Saclay - Ile de France) – Université Paris XI - Paris Sud, CNRS : UMR8623, Polytechnique - X, INRIA – Bât. Alan Turing ; Campus de l'Ecole Polytechnique ; 1 rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France
[4] LORIA – CNRS : UMR7503 – France

INTRODUCTION

Protein-RNA complexes participate in many aspects of cell regulation, and their atomistic structural description is crucial to understand, modulate or engineer the recognition mechanism. As the experimental resolution of their structure is arduous, computational protein-RNA docking methods have been developed, that aim at modeling a 3D assembly by assembling structures of each isolated constituent. Yet for highly flexible objects like single-stranded RNA (ssRNA), the isolated structure of the whole molecule can adopt an ensemble of conformations too large to be experimentaly solved or computationaly modeled. Therefore, if somehow successful on structured RNAs, classical computational docking methods cannot handle the flexibility of ssRNA.

We have recently proposed an original fragment-based approach to accurately model ssRNA-protein complexes from protein structure and RNA sequence, consisting in (i) cutting the RNA sequence into trinucleotides overlapping by 2 nucleotides, represented by a fragment library built from known protein-RNA structures; (ii) docking each conformer-ensemble separately onto the protein; (iii) assembling the spatially compatible poses into an RNA chain.

This method can either blindly determine the RNA-binding site with high specificity [de_Beauchene, de_Vries, Zacharias, PloS 2016], or model the full RNA based on protein-RNA contacts predicted by homology [de_Beauchene, de_Vries, Zacharias, NAR 2016]. In the absence of known contacts, the number of compatible fragment chains is beyond the reach of brute force approaches.

Here, we present an improved method capable of modeling the full bound ssRNA without homology information. Improvements include:

i. a new docking protocol for sampling deep binding pockets;

ii. a stochastic backtracking algorithm for unbiased sampling of chains from the fragment connectivity graph, after computing the partition function of each pose by dynamic programming;

*Speaker

iii. a combination of filters based on biophysical characteristics of the binding site.

As a proof-of-principle, we successfully applied this method on a poly-U ssRNA inserted in the deep cavity of an exonuclease. The accuracy of 4 Å RMSD reached for this 10-mer ssRNA is far beyond the reach of any other docking program.

RESULTS

To evaluate the quality of our docking results, we computed the RMSD (Root Mean Squared Deviation) of the fragment poses or the total RNA chains to their reference position in the crystallographic structure. Given the very high flexibility and the hybrid size of our ssRNA , we adapted the classical acceptance criteria of 2 and 10 Å used for macromolecules and small-ligand docking, toward 3 and 5 Å for fragments and 8-mer chains respectively.

1. A new docking protocol for buried binding

For the fragment assembly to succeed, the sampling of each individual fragment from the sequence needs to be sufficiently comprehensive. We formerly used the ATTRACT docking software [Zacharias, Proteins 2003], which performs a minimization of the intermolecular energy in an empirical force field, starting from random positions of a ligand around a receptor. As the ligand can quickly be trapped in local minima of the energy landscape at a protein surface, the number of starting positions must be large to increase the probability of the ligand to find the global minimum. Our previous ATTRACT protocol for docking RNA on a globular protein surface started from 3.10ˆ7 positions, among which the 2445 UUU conformers of our fragment library were randomly distributed. The 10ˆ6 best-scored poses were retrieved. Yet for docking inside the deep buried cavity of our exonuclease, this sampling was not large enough. Only a small fraction of the starting positions could enter the cavity without getting stuck at the protein surface, resulting in only 0 – 42 correct poses per fragment, with a best-RMSD up to 3.4 Å.

Therefore, we developed DeepATTRACT, a new protocol for docking inside deep cavities, and compared the sampling quality with the previous ATTRACT protocol. DeepATTRACT uses the detection of pocket points by the POCASA server [Yu, Zhou, Tanaka, Yao, Bioinformatics 2010] and selects as starting positions the points with enough neighors to accommodate a trinucleotide. The number of such points (5682) being too large for all the 2445 UUU conformers to be tested at each point, we used a ”hierarchical sampling”:

i. The library conformers were clustered by RMSD in 108 clusters;

ii. Each cluster center was placed at each starting point with 32 different orientations;

iii. Each combination (point * conformers * orientations) was scored with the ATTRACT function;

iv. When a good score (low energy) was found, each conformer in the same cluster was placed at the same position;

v. The new combinations were shortly minimized and re-scored, and the 10ˆ6 best-scored poses were retrieved

With this new protocol, 54 – 644 acceptable poses were found for each fragment, with a best-

RMSD in range 1.0 – 2.6 Å. As expected, this strong overall improvement over the previous ATTRACT protocol is particularly pronounced for the four most buried fragments: the number of acceptable poses inproved from 0 - 2 to 56 – 644, and the best-RMSD from 2.1 – 3.4 Å to 1.2 – 2.6 Å.

2. Assembly by stochastic backtracking

We then searched chains of compatible docking poses with a low total binding energy. As a first approximation, we used the ATTRACT score as a proxy for the pose binding energy, and additively defined over assemblies. The compatibility criteria between two successive poses was defined as an RMSD of the two overlapping nucleotides below 2 Å. Assembling 10ˆ6 poses per fragment would lead to 10ˆ48 possible 8-fragments chains. To retrieve the most probable assemblies while avoiding a brute-force enumeration, we used a new algorithm for unbiased sampling of chains.

The connectivity of each pair of poses was evaluated, resulting in a directed graph of connected poses. The partition function Z was computed over the set of all admissible assemblies, using dynamic programming. As a side product, the algorithm computes the exact Boltzmann probability of a given pose to participate in a downstream path (cf Methods). It can then be adapted to perform a stochastic sampling of the Boltzmann ensemble, resulting in a good approximation of the Boltzmann ensemble of low-energy. We iterated our sampling procedure and obtained 10ˆ5 chains.

By repeating this sampling, we obtained 10ˆ5 chains. After averaging the coordinates of overlapping nucleotides, we obtained a best RNA at 2.2 Å RMSD from the reference structure. But the fraction of acceptable models was low (3), and the scoring function of ATTRACT is not precise enough to select the best models, requiring the use of more effective filters.

3. Enrichment by combinable filters

To enrich the fraction of correct models, we used general and system-specific knowledge to define geometric constraints as filters:

(a) Mg2+ ions are well-known for being chelated by RNA phosphate groups and to stabilize RNA-protein complexes. One such ion is present at the bottom of the exonuclease pocket. We imposed the last RNA phosphate of our chain to be within 7 Å from it.

(b) Aromatic rings at protein-RNA interfaces are well-known for establishing stacking interactions with RNA bases. One aromatic ring is present at the entrance of the exonuclease binding pocket of our exonuclease. Based on the pocket size and the average nucleotide size, we imposed the 1st base of our 10-mer RNA chain to be within 5 Å from the aromatic ring.

(c) We assume a linear ssRNA , i.e. nucleotides from distinct fragment at more than 6 Å from each other.

Applying each filter separately retained pools with 0.5 – 12% correct models (enrichment x1.6 – x34). Combining two filters retained pools with 12 – 50% correct models (x2.7 – x163). The most effective filters were (b), (a) then (c), (c) being mosty redundant with (b). Finally, combining the three filters retained only one model, with an RMSD of 4.0 Å.

The complete process took few CPU hours.

DISCUSSION AND PERSPECTIVES

We present a method capable of modeling a protein-bound ssRNA based on the protein structure and ssRNA sequence. First, our new docking protocol for docking RNA fragments inside deep pockets improved the sampling quality for all buried fragments. Second, a new stochastic backtracking algorithm to perform unbiased sampling from a connectivity graph of the docked fragments generated near-native RNA chains among 100,000 samples. Finally, an efficient and effective filtering procedure to incorporate knowledge on the protein-ssRNA system led to a fraction in correct models of up to 50–100% after applying 2-3 filters. As a first proof-of-principle, the method could model *ab initio* a 10-mer bound ssRNA, an unprecedented length far beyond the reach of standard small-molecule or macromolecular docking programs.

However, several limitations must be regarded:

First, filters such as those used on that particular case are not always available solely from the knowledge of the protein structure. Additional experimental data (mutagenesis, cross-linking...) can be required. The advantage of this sample-then-filter approach compared to a constrained sampling is to be able to predict, from the initial sample, which set of experiments would best partition it. This reduces the number of experiments required for a given targeted enrichment factor.

Second, the bound structure of the protein was here considered as exactly known, while in a real-case docking, only an unbound structure of the protein can be known. Conformational changes between the bound and unbound protein are then likely to diminish the accuracy of the results.

Third, we considered the single-stranded state of the bound RNA as known a priori. In some cases, thisinformation is not available and must be retrieved from the modeling.

We consider several ways of overcoming these limits in our future research:

Regarding the scoring limitation, we have so far neglected: (i) the scoring specificities of fragments over full molecules, (ii) the internal energy of our fragments, and the compensation effect of e.g. breaking intra-fragment base stacking to permit stacking with protein residues, and (iii) the quality of pair connectivity, by using a simple boolean criterion. To increase the fraction of correct models in our initial sampling, we will (i) parametrize a new fragment-specific function to estimate the binding energy, (ii) sum of the internal and binding energies of each pose, (iii) weight the edges of the graph by the connectivity quality.

Regarding protein flexibility, we will use homology models with more or less divergence, MD or unbound forms (if available) in order to investigate its impact on the sampling quality of poses. If necessary, to model a flexible protein while avoiding the enumeration of its possible conformations, we will apply the same principle of decoupled sampling as for the RNA, but in a hierarchical way, by representing the protein as a tree of global and local conformations. Each ”local” set of conformations can be used for docking, then the compatibility of conformations can be assessed together with the connectivity of the RNA poses bound to them.

To generalize the method to RNA of arbitrary base-pairing (secondary structure, ”2D”), we will create a new 2D-specific fragment library, use a sequence-based prediction of a Boltzmann ensemble of 2D structures, dock all fragments of possible 2D structure at each sequence position, then incorporate the 2D probability in the assembly.

METHODS SUPPLEMENTARY

DeepATTRACT: As we dealt with a poly-U, one single docking of a UUU fragment was performed, and the poses were compared to each fragment at each position in the reference structure. POCASA was used with a 2 Å probe and a 1 Å spacing grid. Points with more than 500 neighbors within 7Å were retrieved. The fragment library was clustered with a 3 Å RMSD cutoff to keep representatives. Poses with an ATTRACT score below 100 kcal/Mol were kept for further testing with the whole corresponding cluster. The final poses were clustered with a 2 Å cutoff.
Partition function Z: For a pose i at position k, Z(k, i) is the sum, for all j connected to i, of {exp[(E(i) + E(j)) / RT] times Z(k-1, j)}, where E(i) is the ATTRACT score of pose i, and RT the Bolztman factor. In the stochastic backtracking, each pose is chosen with a probability Z(k,i) / sum(Z(k,j) over j).

Keywords: RNA modeling, RNA 3D structure, RNA, protein docking, combinatorial assembly

Comparing protein structures with RINspector automation in Cytoscape

Guillaume Brysbaert* [1], Théo Mauri † [1], Marc Lensink [1]

[1] UMR8576 (UGSF) – Centre National de la Recherche Scientifique - CNRS, Université Lille Nord (France) – France

In biology, understanding and knowing a protein function is very important but is an intricate task. Thus, integrated approaches like *in silico* methods based on sequences and structures could solve this problem. A PDB file contains all the information about a structure and can be used directly for many analyses or transformed for other types of analyses. That is the case of Residue Interaction Networks (RINs). These networks are built from protein structures where residues are nodes and the edges are detected interactions between these residues. Analyses performed on these networks like centrality analyses permit to give clues of the involvement of key residues in the function or the folding of a protein (1,2).
To complement the network analyses, the backbone flexibility and changes upon mutations can be computed to help in the search for key residues in the design of mutagenesis experiments and to unravel the function of a protein (3,4).
RINspector is an app we developed for the Cytoscape network analysis and visualization software (5) to analyze residue interaction networks and visualize flexibility predictions associated to the protein sequence. This app allows the user to make centrality analyses based on shortest path lengths with associated Z-scores and to analyze a protein flexibility by querying the DynaMine server (6,7). The results can be display on a graph of interactive flexibility which permits to select residues to mutate in order to compare the new graph with the wild type. A connection between the RIN, the flexibility graph and the structure (with structureViz app/Chimera if installed (8,9)) selects a residue in the three representations and see it in its context. This tool allows to make a quick identification of key residues in protein function and stability.

The RINspector app(10) is useful and convenient if the networks are not too large and numerous but may be very greedy in memory and CPU in certain kind of centrality analyses or if the amount of RINs/nodes raises. An example is the analysis of conformers from a NMR experiment or the comparaison of RINs generated from a wild type structure and mutants. Thanks to the CyREST technology(11) which is implemented in Cytoscape, RINspector now embeds a documented API that provides automation of flexibility predictions and centrality calculations. This enables users to complement analyses done in this software with scripts which can be written in external languages like Python or R.

We present two examples which take benefit of the API and automatically present recap charts. The first one considers NMR data with 10 conformers of a yeast N-acetyleglucosamine transferase, all grouped in one PDB ID. The second one is the TetratricoPeptide Repeat (TPR)

*Corresponding author: guillaume.brysbaert@univ-lille1.fr

†Speaker

domain of a human O-GlcNAc Transferase (OGT) which is known to play a role in the substrate recognition of this enzyme (12). In the latter, five residues were mutated to see their impact on centralities in the RINs and on the domain flexibility.

RINspector is available in the Cytoscape app store and published in Bioinformatics (10).

References:

1. del Sol A, Fujihashi H, Amoros D, Nussinov R. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol Syst Biol. 2006;2:2006.0019.

2. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, et al. Network analysis of protein structures identifies functional residues. J Mol Biol. 2004 Dec 3;344:1135–46.

3. Teague SJ. Implications of protein flexibility for drug discovery. Nat Rev Drug Discov. 2003 Jul;2:527–41.

4. Golovanov AP, Hawkins D, Barsukov I, Badii R, Bokoch GM, Lian LY, et al. Structural consequences of site-directed mutagenesis in flexible protein domains: NMR characterization of the L(55,56)S mutant of RhoGDI. Eur J Biochem. 2001 Apr;268:2253–60.

5. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13:2498–504.

6. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF. From protein sequence to dynamics and disorder with DynaMine. Nat Commun. 2013;4:2741.

7. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF. The DynaMine webserver: predicting protein dynamics from sequence. Nucleic Acids Res. 2014 Jul;42:W264-70.

8. Morris JH, Huang CC, Babbitt PC, Ferrin TE. structureViz: linking Cytoscape and UCSF Chimera. Bioinformatics. 2007 Sep 1;23:2345–7.

9. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct;25:1605–12.

10. Brysbaert G, Lorgouilloux K, Vranken W, Lensink MF. RINspector: a Cytoscape app for centrality analyses and DynaMine flexibility prediction. Bioinforma Oxf Engl. 2017 Sep 22;

11. Ono K, Muetze T, Kolishovski G, Shannon P, Demchak B. CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API. F1000Research [Internet]. 2015 Aug 5 [cited 2018 Apr 20];4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4670004/
12. Rafie K, Raimi O, Ferenbach AT, Borodkin VS, Kapuria V, van Aalten DMF. Recognition of a glycosylation substrate by the O-GlcNAc transferase TPR repeats. Open Biol [Internet]. 2017 Jun 28 [cited 2018 Apr 20];7(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5493779/

Keywords: rinspector, cytoscape, protein structure, residue interaction network, centrality analysis,

flexibility prediction, automation, structure ensemble

Evaluation of the docking performance of open source algorithms on carbohydrate binding domains

Charlotte Brunel-Cadic * [1], Cyrille Grandjean [2], Stéphane Téletchéa [1]

[1] Unité de Fonctionnalité et Ingénierie des Protéines (UFIP) – Université de Nantes, CNRS : UMR6286 – 2 rue de la Houssinière Bâtiment 25 44322 Nantes cedex 3, France
[2] Unité Fonctionnalité et Ingénierie des Protéines (UFIP-université de Nantes) – UFIP Nantes – 2 rue de la houssinière 44322 Nantes cedex 3, France

This study consists in studying the binding of (mono-)saccharides to galectins, and to assess the performance of docking algorithms on these small molecules. Carbohydrates are small hydroxylated molecules with an important intrinsic flexibility, they also possess a very important hydrophobic patch largely ignored by classical molecular mechanics applied in docking algorithms (1,2). Altogether this lack of precise representation of this small category of biological constituents need to be quantified. Our objective is to compare the ability of open source docking software to predict the correct pose. Since many human galectin binding modes are well documented, we started our analysis using re-docking strategies. We also set up a complex analysis protocol to assess the accuracy and the precision of each algorithm scoring function. We shall present the results of our analysis and the recommendations for general-purpose enhancement of existing algorithms or scoring functions.

Bibliography

1. Atmanene C, Ronin C, Téletchéa S, Gautier FM, Djedaíni-Pilard F, Ciesielski F, Vivat V, Grandjean C. Biophysical and structural characterization of mono/di-arylated lactosamine derivatives interaction with human galectin-3. Biochem Biophys Res Commun. 2017;489(3):281-286

2. Stanca-Kaposta EC, Gamblin DP, Screen J, Liu B, Snoek LC, Davis BG, Simons JP. Carbohydrate molecular recognition: a spectroscopic investigation of carbohydrate-aromatic interactions. Phys Chem Chem Phys. 2007;9(32):4444-51

Keywords: molecular docking

*Speaker

Repository of Enriched Structures of Proteins Involved in the Red blood Cell Environment (RESPIRE)

Stéphane Téletchéa * [1], Hubert Santuz , Sylvain Leonard , Catherine Etchebest [2]

[1] Unité de Fonctionnalité et Ingénierie des Protéines (UFIP) – Université de Nantes, CNRS : UMR6286 – 2 rue de la Houssinière Bâtiment 25 44322 Nantes cedex 3, France
[2] Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB) – Université Paris Diderot, Sorbonne Paris Cité, INTS, Inserm : UMRS1134 – 6 rue Alexandre Cabanel, 75739 Paris cedex 15, France

The Red Blood Cell (RBC) is a metabolically-driven cell vital for processes such a gas transport and homeostasis. RBC possesses at its surface exposing antigens proteins that are critical in blood transfusion. Due to their importance, numerous studies address the cell function as a whole but more and more details of RBC structure and protein content are now studied using massive state-of-the art characterisation techniques. Yet, the resulting information is frequently scattered in many scientific articles, in many databases and specialized web servers. To provide a more compendious view of erythrocytes and of their protein content, we developed a dedicated database called RESPIRE that aims at gathering a comprehensive and coherent ensemble of information and data about proteins in RBC. This cell-driven database lists proteins found in erythrocytes. For a given protein entry, initial data are processed from external portals and enriched by using state-of-the-art bioinformatics methods. As structural information is extremely useful to understand protein function and predict the impact of mutations, a strong effort has been put on the prediction of protein structures with a special treatment for membrane proteins. Browsing the database is available through text search for reference gene names or protein identifiers, through pre-defined queries or via hyperlinks. The RESPIRE database provides valuable information and unique annotations that should be useful to a wide audience of biologists, clinicians and structural biologists.

Database URL: http://www.dsimb.inserm.fr/respire/

Keywords: Red Blood Cell proteins, Protein Structure Prediction, Membrane Proteins

*Speaker

Could MAPKs watch in the mirror? Exploring large-scale phosphoproteomics data to assess whether MAPKs could phosphorylate the Pro-(Ser/Thr) motif in addition to the canonical (Ser/Thr)-Pro motif.

Emmanuelle Lastrucci *† [1,2], Jean Bigeard [3,4], Naganand Rayapuram [5], Ronny Volz [5], Hanna Alhoraibi [5], A.aala Abulfaraj [5], F. Javier Guzman [6], Afaque Ahmad Momin [6], Stefan T Arold [6], Heribert Hirt [5], Delphine Pflieger‡ [1,2]

[1] Centre National de la Recherche Scientifique, EDyP/BGE/BIG (CNRS) – CNRS : FR3425 – 17 rue des martyrs, 38000, Grenoble, France

[2] Commissariat à l´nergie atomique et aux énergies alternatives (CEA) – Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble : BIG/BGE, , Universitey of Grenoble Alpes (UGA), INSERM U1038 – grenoble, France

[3] Institute of Plant Sciences Paris-Saclay IPS2, CNRS, INRA, Université Paris-Sud, Université Evry, Université Paris-Saclay – Université Paris-Sud - Université Paris-Saclay – France

[4] Institute of Plant Sciences Paris-Saclay IPS2, Paris Diderot, Sorbonne Paris-Cité – Université Sorbonne Paris Cité (USPC) – France

[5] Center for Desert Agriculture, 4700 King Abdullah University of Science and Technology (KAUST) – Thuwal, Saudi Arabia

[6] Division of Biological and Environmental Sciences and Engineering (BESE),Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST) – Thuwal, Saudi Arabia

Phosphorylation is one of the most important and extensively studied post-translational modifications (PTM) in proteins. It consists of the attachment of a phosphoryl group mainly occurring on serine, threonine and tyrosine residues in eukaryotes. The addition of the phosphoryl group can change the protein conformation and thus the properties of proteins such as their enzyme activity, subcellular localization, their stability and interaction with other proteins [1]. Protein phosphorylation is a reversible PTM mediated by protein kinases, enzymes that phosphorylate, and protein phosphatases, enzymes that de-phosphorylate proteins. Protein kinases are among the largest gene families in eukaryotes [2] and are classified into subfamilies according to the specific amino acid residues that they phosphorylate. Most protein kinases act on both serine and threonine residues (serine/threonine kinases), others act on tyrosine residues (tyrosine kinases) and a number act on all three (dual-specificity kinases) [3,4]. In addition

*Speaker

†Corresponding author: Emmanuelle.LASTRUCCI@cea.fr

‡Corresponding author: Delphine.PFLIEGER@cea.fr

to this classification, protein kinase substrates are also categorized into different subgroups according to a consensus sequence which corresponds to the amino acid sequence surrounding the phosphorylated Ser / Thr / Tyr residues [4,5]. Despite the wealth of studies in this area, even for well-known kinases, new motifs can be discovered, as shown in [6] which recently discovered a new motif for the cyclin B-dependent protein kinase Cdk1.
One kinase family corresponds to the mitogen-activated protein kinases (MAPKs) which regulate various cell functions [7,8]. This kinase family is specific to Ser and Thr residues and is associated with the low-stringency motif (S/T)*P (where * represents the phosphorylation) and the high-stringency motif PX(S/T)*P (where X represents any amino acid). In *Arabidopsis thaliana*, the MAPKs MPK3, MPK4 and MPK6 have been implicated in the regulation of cell cycle, cytokinesis, plant development and innate immunity [9]. While some substrates of MPK3, MPK4 and MPK6 have been identified, the picture is still far from complete. In order to identify substrates of these immune MAPKs in *Arabidopsis thaliana*, we performed an *in vivo* large-scale phosphoproteomic analysis of wild-type and of *mpk3*, *mpk4*, and *mpk6* deletion mutants following treatment with a microbe-associated molecular pattern (MAMP), the 22-amino-acid long peptide flg22 which corresponds to a well-conserved sequence of bacterial flagella [10]. To do so, we used a combination of liquid chromatography and tandem mass spectrometry (LC-MS/MS). In MS/MS analysis, peptides are most commonly identified by matching fragmentation spectra against theoretical spectra of all candidate peptides represented in a generalist protein sequence reference database [11].

This study allowed the identification of 70 peptides that harboured an (S/T)*P motif and whose phosphosite abundance was significantly affected by deletion of any one of the MAPKs and/or by flg22 treatment [10]. As the peptides contained the low-stringency motif of MAPKs, we hypothesized that at least some of these substrates were phosphorylated by MAPKs. In order to validate some of them, *in vitro* kinase assays followed by mass spectrometry analysis (KA-MS) were performed on 11 candidate proteins in the presence of either MPK3, MPK4 or MPK6. A kinase assay (KA) consists in incubating a protein with a specific kinase to see if the protein can be phosphorylated by this kinase. These assays validated the selected candidates, but intriguingly, upon incubation with MPK3 or MPK6, we also identified a few peptide sequences exhibiting PS* or PT* sites, which we will call ”MAPK mirror motif”, whereas the (S/T)*P motif will be called ”usual”. Following this observation and given the recent discovery of new consensus motifs for well-known kinases [6], we wanted to assess if MAPKs are able to phosphorylate protein substrates in the PS*/PT* context and if this phosphorylation could be biologically relevant. To answer this question, we took advantage of publicly available data corresponding to *in vitro* and *in vivo* assays made on plant and mammals. The methodologies and results are presented below.

In order to assess whether these mirror sites can be phosphorylated by MAPKs and might be biologically relevant, we first estimated the stoichiometry of modification at these sites, compared to the stoichiometry observed at usual sites. For this purpose, we used the *in vitro* KA-MS experiments carried out on the 11 selected candidate proteins in the presence of either MPK3, MPK4 or MPK6. The level of phosphorylation at all sites was calculated as the ratio of the abundance of the phosphorylated peptide over the total abundance of the peptide (phosphorylated peptide plus the corresponding non-modified peptide). The raw data file acquired on the 33 samples were converted into peak lists using the program Mascot Distiller and searched against the *A. thaliana* database using Mascot. This step tries to match the fragments ion experimental MS/MS spectra with the MS/MS spectra generated *in silico* on tryptic peptides from proteins recorded in the database (taking into account all possible modifications per amino acid) to provide a list of tentatively identified peptides. The Mascot results were then imported into Proline [12], a software suite dedicated to proteomics data developed in the lab (http://www.profiproteomics.fr/proline/), in order to obtain only confident peptides (Mascot score $> =25$) and their quantitative abundances. To estimate the stoichiometry, we only considered phosphopeptides with confident phosphorylation sites, i.e. single phosphorylated peptides with a site confidence (probability that the phosphorylation is present at this specific residue) superior to 0.8. The 33 KA-MS experiments allowed to identify 104 phosphopeptides of which 58 were considered to bear a reliable phosphorylation site. We could estimate the stoichiometry of 48 phosphopeptides (due to the non-detection of the corresponding sequence in a non-modified form for a few phosphopeptides). Among the 48 phosphopeptides, 36 had a usual low stringency (S/T)*P motif and five exhibited the mirror motif P(S/T)*. Even if the mean stoichiometry obtained for mirror sites was less than for usual sites, 9% and 25% respectively, they nevertheless compared very well with the stoichiometry observed for phosphopeptide AT1G11360 that we determined to be a shared *in vivo* target of MPK4 and MPK6. Therefore, we verified that MPK3 and MPK6 could phosphorylate P(S/T)* sites *in vitro* on a few putative substrates and the hypothesis that the observed mirror motifs might have a biological meaning could not be ruled out by the rather low phosphorylation levels observed in these KA-MS.

The above kinase assays allowed assigning to MPK3 and MPK6 the *in vitro* ability to phosphorylate mirror sites. However, kinase assay can force phosphorylation due to the lack of complexity. We next wanted to assess to what extent such phosphorylation at mirror P(S/T) sites might be observed *in vivo* in whole plants. We then explored the large-scale phosphoproteomics datasets that we acquired on cytoplasmic proteins extracted from WT, *mpk3*, *mpk4* and *mpk6* plants that had been either flg22-or mock-treated [10]. With the goal to start with the most confident phosphorylation sites, we developed a methodology which applies a more stringent filtering than the original data. Briefly, we first asked for a Mascot score and a Mascot delta score (MD-score) which allowed to reach a 99% confidence on phosphosite assignments according to [13]. A significant number of phosphorylated sequences contained symmetrical stretches PTTP. These peptides constituted particularly difficult cases in which to localize the phosphosite, and as a consequence were hard to confidently attribute to the category of usual (PTT*P) or of mirror (PT*TP) sites. We then defined a range of six different cases to classify the motifs: "otherKinases", "ambiguous", "mirrorWeak", "usualWeak", "usual" and "mirror". Last, we also filtered out the sequences phosphorylated at sites that could match both the MAPK motif and other kinases in *A. thaliana*. From the 16,067 initial phosphopeptides identified in [10], 8 388 passed the above filters, which corresponded to 838 unique phosphopeptides and 1003 phosphosites. Finally, 17 phosphopeptides corresponding to 18 distinct proteins were identified in this analysis with the mirror motif. About half of the above PS*-containing peptides showed a decreased abundance in the *mpk6* mutant, which supports the hypothesis that they are phosphorylated by this MAPK *in planta*.

Last, we wanted to assess whether a mirror site might also be observed with mammalian MAPKs. We then explored the phosphoproteomics data produced in the cytosolic and nuclear fractions of rat cells with the goal to identify substrates of ERK1/2, MAPKs conserved in mammals [14]. The methodology used in this analysis to obtain only confident phosphopeptide identification and localization was similar to the *in vivo* A. *thaliana* study except for a few steps that we had to adapt to the different input files. As the localization site confidence was not available in the original data, we calculated it using the Proline software suite [12] and kept peptides for which the phosphorylation probability was above 0.8. We did not filter out sites matching other mammalian kinases because a specific inhibitor of ERK1/2 was used, allowing to extract only peptides phosphorylated by these two MAPKs. From the 3015 and 5222 unique phosphopeptides identified in the cytosolic and nuclear fractions, 1612 and 2777 passed the Mascot score threshold of 25 and a localization site confidence of 0.8, respectively. From these, 548 and 1204 corresponded to the usual motif and 33 and 56 corresponded to mirror motifs. Finally, a few could be considered to be ERK1/2-dependent: 39 and 105 for usual sites and 2 and 7 for mirror sites.

The goal of this study was to assess if the MAPK usual consensus motif (S/T)*P could be extended with the additional mirror motif P(S/T)*. For this purpose, we decided to take advantage of publicly available data and develop methodologies to extract only confident phosphopeptides containing this mirror site. Unfortunately, our results do not allow answering this question in a statistical way. *In silico* structure modelling is currently underway to see whether the crystal structure of ERK may accept for phosphorylation, in its catalytic loop, P(S/T)-containing sequences identified above in mammals and in *Arabidopsis*.

1. Bigeard Jean, Rayapuram Naganand, Pflieger Delphine, Hirt Heribert. Phosphorylation-dependent regulation of plant chromatin and chromatin-associated proteins. PROTEOMICS. 2014;14:2127–40.

2. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The Protein Kinase Complement of the Human Genome. Science. 2002;298:1912–34.

3. Ardito F, Giuliani M, Perrone D, Troiano G, Muzio LL. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). Int J Mol Med. 2017;40:271–80.

4. Kemp BE, Pearson RB. Protein kinase recognition sequence motifs. Trends Biochem Sci. 1990;15:342–6.

5. Amanchy R, Periaswamy B, Mathivanan S, Reddy R, Tattikota SG, Pandey A. A curated compendium of phosphorylation motifs. Nat Biotechnol. 2007;25:285–6.

6. Suzuki K, Sako K, Akiyama K, Isoda M, Senoo C, Nakajo N, et al. Identification of non-Ser/Thr-Pro consensus motifs for Cdk1 and their roles in mitotic regulation of C2H2 zinc finger proteins and Ect2. Sci Rep [Internet]. 2015 [cited 2018 Apr 17];5. Available from: http://www.nature.com/articles/srep07929

7. Chen Z, Gibson TB, Robinson F, Silvestro L, Pearson G, Xu B, et al. MAP Kinases. Chem Rev. 2001;101:2449–76.

8. Platanias LC. Map kinase signaling pathways and hematologic malignancies. Blood. 2003;101:4667–79.

9. Rodriguez MCS, Petersen M, Mundy J. Mitogen-Activated Protein Kinase Signaling in Plants. Annu Rev Plant Biol. 2010;61:621–49.

10. Rayapuram N, Bigeard J, Alhoraibi H, Bonhomme L, Hesse A-M, Vinh J, et al. Quantitative Phosphoproteomic Analysis Reveals Shared and Specific Targets of Arabidopsis Mitogen-Activated Protein Kinases (MAPKs) MPK3, MPK4, and MPK6. Mol Cell Proteomics. 2018;17:61–80.

11. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics. 2010;73:2092–123.

12. Proline Main Page [Internet]. ProFi Proteomics. [cited 2018 Apr 19]. Available from: http://www.profiproteomics.fr/proline/

13. Savitski MM, Lemeer S, Boesche M, Lang M, Mathieson T, Bantscheff M, et al. Confident Phosphorylation Site Localization Using the Mascot Delta Score. Mol Cell Proteomics. 2011;10:M110.003830.

14. Courcelles M, Frémin C, Voisin L, Lemieux S, Meloche S, Thibault P. Phosphoproteome dynamics reveal novel ERK1/2 MAP kinase substrates with broad spectrum of functions. Mol Syst Biol. 2013;9:669.

Keywords: Proteomic, Phosphorylation, Mass spectrometry, Kinases, MAPKs, consensus motif, kinase assay

In silico protein digestion

Nicolas Maillet * [1]

[1] Hub Bioinformatique et Biostatistique - Bioinformatics and Biostatistics HUB – Institut Pasteur [Paris], Centre National de la Recherche Scientifique : USR3756 – C3BI, 25-28 rue du Docteur Roux, 75724 Paris cedex 15, France

Proteases, also known as proteolytic enzymes, have been studied for more than 80 years. (1) Those enzymes are widely used in industry, medicine and as a biological research tool, for example in protein characterization or more generally in proteomics and proteogenomics. (2)

Recently, interest in proteases has gained importance due to advancements in mass spectrometry techniques used in proteomics and proteogenomics. In "bottom-up" analysis, using tandem mass spectrometry (MS/MS), optimal peptide size range is 600–5,000Da (3) when proteins size are usually more than 10000 Da. Therefore, for bottom-up approaches, protein digestions are required. To perform digestions, one or several proteases, like trypsin, pepsin or thrombin, are used. Each protease has specific cleavage sites relying on solvent accessibility, pH, temperature, etc. The use of different proteases individually or in combination creates a unique set of peptides. Performing multiple digestions can increase overall confidence in protein identification if cleaving sites are different. It is not always easy to determine which combination of enzymes will lead to a set of peptides suitable for MS/MS analysis. However, the cost of some enzymes does not allow for easily trying several combinations to avoid redundancy of cleaving sites. Few software exist that predict cleavage sites of proteases in protein sequences. Among those, the most commonly used are PeptideCutter from ExPASy Server (4) and a module integrated in MaxQuant. (5)

PeptideCutter performs a digestion using one or several enzymes, among a total list of 38, and provides detailed results, including positions of cleavage site, peptide sequences, length and mass. Despite the valuable information provided by the specific software, three main features are missing.
First, in order to thoroughly analyze the behavior of a specific combination of enzymes, it is important to try this combination on many different proteins. With PeptideCutter the user cannot perform parallel or automatic sequential digestions of more than one sequence and thus this procedure is time consuming and not efficient.
The second drawback of this tool is how a combination of enzymes is used. In PeptideCutter, all selected enzymes are supposed to be present at the same time during digestion. It is therefore difficult to simulate distinct digestions, i.e. digestions of the same sequence using different enzymes separately. This means that instead of an automatic succession of distinct digestions, one has to run the software as many times as the number of distinct digestions, multiplied by the number of concerned sequences.
Last but not least, in PeptideCutter it is not possible to input novel enzyme definitions. As

*Speaker

previously mentioned, there is a growing interest in proteases and new or more specific enzymes (denoted as "Sequencing Grade", or SG) are developed. Depending on the company manufacturing those SG enzymes, specificity and definition can change. Hence, it is important for a user to easily adapt the software by including novel definitions of enzymes.

MaxQuant partially overcomes those issues. The user can input new enzyme definitions by specifying between which amino acids cleavages occur. Unfortunately, this definition is not sufficient to properly define some enzymes. For example, definition of Trypsin in MaxQuant lacks some exceptions. It is defined as cleaving after K or R, but not before P. However, it has been reported (6) that although most of the times P blocks the cleavage when found after K, this is not true when K is preceded by W: a cleavage occurs after K in 'WKP' motif. Currently, it is not possible to create such rules in MaxQuant.

This talk presents Rapid Peptides Generator (RPG), a new standalone software dedicated to predict proteases-induced cleavage sites on sequences. RPG is a python tool taking (multi-)fasta/fastq file of proteins as input and digest each of them. Digestion mode can be either 'concurrent', i.e. all enzymes are present at the same time during digestion, or 'sequential'. In sequential mode, each protein will be digested by each enzyme, one by one. The resulting peptides contain the same informations as PeptideCutter, as-well as an estimation of isoelectric point (pI) of each peptide. Shortly, the isoelectric point is the pH at which a peptide carries no net electrical charge and a good approximation can be computed on small molecules. Results are outputted in multi-fasta, CSV or TSV file. Currently, 42 enzymes and chemicals are included in RPG. User can easily design new enzymes, using a simple yet powerful grammar. This grammar allows the user to design complex enzymes like trypsin or thrombin, including many exceptions and different cleavage sites. User-defined enzymes are then interpreted by RPG and included in the local installation of the software. RPG has been developed in a way to reproduce exactly the cleaving results of PeptideCutter, with the exception of enzymes where PeptideCutter can not be as specific as RPG.

RPG is available through pip ('pip install rpg') and follows the standards for software development with continuous integration on Gitlab (https://gitlab.pasteur.fr/nmaillet/rpg) and automatic on-line documentation (https://rapid-peptide-generator.readthedocs.io).

1. Neurath, H. Proteolytic enzymes, past and future. Proc Natl Acad Sci USA 96, 10962–10963 (1999).
2. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. Nature Publishing Group 11, 1114–1125 (2014).
3. Engel, L., Saveliev, S., Urh, M., Simpson, D., Jones, R. and Wood, K. Using Endoproteinases Asp-N and Glu-C to Improve Protein Characterization. [Internet] . [cited: 2018, 04, 20]. Available from: http://france.promega.com/resources/pubhub/using-endoproteinases-asp-n-and-glu-c-to-improve-protein-characterization/
4. Gasteiger E. et al. (2005) Protein Identification and Analysis Tools on the ExPASy Server. In: Walker, J. M. The Proteomics Protocols Handbook. (Springer Science & Business Media, 2007).
5. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology 26, 1367–1372 (2008)
6. Keil, B. Specificity of Proteolysis. (Springer Science & Business Media, 2012). doi:10.1007/978-3-642-48380-6

Keywords: protein, peptide, enzyme, protease, digestion

ARIAEC: protein structure folding from evolutionary couplings and NMR restraints

Fabrice Allain *† 1, Benjamin Bardiaux‡ 1, Mickael Nilges§ 1

[1] Unite de Bioinformatique Structurale – Institut Pasteur de Paris, CNRS : UMR3528 – 28 rue du docteur Roux 75724 Paris, France

In an era of cost-effective genome sequencing technologies, explaining the path from gene expression to phenotypes is a complex challenge. In this scope, structural information plays a key role for the description and understanding of cellular or pathogenic mechanisms [1]. Mapping the mutations onto the related protein structures may for example give insights into the effect of genetic variants and potentially open the path to targeted drug development.

Despite tremendous progress in structure determination from experimental data and the development of fully automated pipelines in large scale structural genomic projects [2], the actual gold standard remains costly and time consuming. As an alternative or complementary solution, computational prediction methods may alleviate the problem by giving preliminary insights of the native structure from solely the amino acid sequence [3]. In the last decades, the field has matured to the point where reliable models can be proposed with or without experimental information to fill this gap [4]. Models based on homologous templates are the most useful form of modeling, but ab initio algorithms remains essential when there is no detectable template. Among existing template free modeling tools, the increasing availability of genomic information from next generation sequencing technologies has brought back the possibility to use evolutionary information in the folding process [5]. The rationale behind this prediction methodology is the following: to maintain energetically favorable interactions and function, amino acid residues in spatial proximity constrain the evolutionary trajectory across the same protein family leading to a network of compensatory mutations. Several approaches such as EVFold [6], pconsFold [7] or CONFOLD [8] combining evolutionary contacts (EC) with structure calculation protocols have been able to generate promising models with various level of success in the last rounds of CASP [9].

Some important aspects of these techniques needs to be improved, including the detection of false positive among the predicted contacts or distinction between intra-subunit and inter-monomeric contacts in multi-subunit proteins. Here the similarity between the types of information provided by evolutionary contacts and by Nuclear Magnetic Resonance (NMR) spectroscopy is striking. NMR experimentations also detects pairs of atoms close in space in order to determine 3D structure of a protein. In the frame of NMR, efficient and robust approaches are available to analyze, assign and apply distance restraints. Among those, the ARIA (Ambiguous Restraints

*Speaker

†Corresponding author: fabrice.allain@pasteur.fr

‡Corresponding author: benjamin.bardiaux@pasteur.fr

§Corresponding author: mickael.nilges@pasteur.fr

for Iterative Assignment) software [10] is one of the most efficient tools to automatically determine structures from NMR data. In this context, our work aims at extending the concept to template free modeling from evolutionary contacts with ARIA.
The work presented here confirms the efficiency of the ARIA protocol for de novo protein structure prediction with EC and the possibility to combine this information with sparse NMR data, validated on two datasets of respectively 15 and 8 proteins [11, 12]. We also show an application on domain modeling from the type VI bacterial secretion system validated by cryo-electron microscopy data and briefly describe the ARIA web server currently under development.

Glusman, G., et al.. (2017). Genome medicine, 9(1), 113.

Joachimiak, A. (2009). Current opinion in structural biology, 19(5), 573-584.

Anfinsen, C. B. (1973).. Science, 181(4096), 223-230.

Schwede, T. (2013). Structure, 21(9), 1531-1540.

G´obel, U., et al.. (1994). Proteins: Structure, Function, and Bioinformatics, 18(4), 309-317.

Marks, D. S., et al. (2011). PloS one, 6(12), e28766.

Michel, M., et al (2014). Bioinformatics, 30(17), i482-i488.

Adhikari, B., et al. (2015).. Proteins: Structure, Function, and Bioinformatics, 83(8), 1436-1449.

Moult, J., et al. (2018). Proteins: Structure, Function, and Bioinformatics, 86, 7-15.

Bardiaux, B., et al. (2012). Protein NMR Techniques (pp. 453-483). Humana Press.

Tang, Y.,et al. (2015). Nature methods, 12(8), 751.

Keywords: structure prediction, ab initio, evolutionary contacts, contact map, nuclear magnetic resonance, simulated annealing, template free modeling

MetaFoldScan: a pipeline to scan metagenomes and identify structural homologs using HMM

Sandra Dérozier *† 1, Véronique Martin‡ 1, Jean-Marc Chatel 2, Nalini Rama Rao 2, Valentin Loux 1, Gwenaëlle André-Leroux 1

1 Unité Mathématiques et Informatique Appliquées du Génome à l'Environnement – MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France – France

2 MICrobiologie de l'ALImentation au Service de la Santé humaine (MICALIS) – Institut National de la Recherche Agronomique : UMR1319, AgroParisTech – F-78350 JOUY-EN-JOSAS, France

Metafoldscan aims at developing a user-friendly interface to intensively scan metagenomes to identify structural homologs associated or not to a target protein. To our knowledge, no comprehensive solution that associates flexible and robust browsing of (meta-)genomes within a reasonable computational time is available to date. Metafoldscan has been set up with the core genome of the human gut microbiota that clusters 57 highly prevalent bacteria [Qin *et al.*, 2010 ; Lin *et al.*, 2014]. This first ecological system has been scanned and Metafoldscan has permit the identification of structural homologs of MAM- Microbial Anti-inflammatory Molecule- from the commensal bacterium Faecalibacterium prausnitzii [Quévrain *et al.*, 2015] and the ubiquitous bacterial protein Mfd [Guillemet *et al.*, 2016] - Mutation Frequency Decline. Both proteins associate biological with therapeutic relevance. Metafolscan is now ready to step into the scaling up to the entire gut microbiota.
MetaFoldScan milestones that have been reached are:

1. the standalone commands for a versatile software to detect structural homologs using HMM [S´oding *et al.*, 2005] and accordingly biologically relevant enzymes in ecological system.

2. the in silico suite of tools set up in 1. is now integrated into the user-friendly Galaxy portal. This includes browsing and filtering upstream meta-omics data.

Now the key issues and challenges are:

1. Set-up of filters and scaling up to the 10 millions genes of the gut microbiota.

2. Validation of the hits and discovery of new enzymes with possibly therapeutic functions.

Références

Qin J., *et al.* **MetaHit Consortium Human gut microbial gene catalogue established**

*Speaker

†Corresponding author: sandra.derozier@inra.fr

‡Corresponding author: veronique.martin@inra.fr

by metagenomic sequencing. *Nature* (2010), 464(7285):59-65.

Lin J., *et al.* **MetaHit consortium: An integrated catalog of reference genes in the human gut microbiome**. *Nat Biotechnology* (2014), 32(8):834-41.

Quévrain E., *et al.* **Identification of an anti-inflammatory protein from Faecalibacterium prausnitzii, a commensal bacterium deficient in Crohn's disease**. *Gut* (2015).

Guillemet E., *et al.* **The bacterial repair protein Mfd confers resistance to the host nitric-oxide response**. *Sci. Report* (2016).
S´oding J., *et al.* **The HHpred interactive server for protein detection and structure prediction**. *Nucleic Acids Res.* (2005).

Keywords: protein structure, metagenomic, galaxy

Ligand-guided homology modelling of the GABAb2 subunit of the GABAb receptor

Thibaud Freyd * [1], Dawid Warszycki [2], Stefan Mordalski [2], Andrzej Bojarski [2], Ingebrigt Sylte† [1], Mari Gabrielsen [1]

[1] Department of Medical Biology, Faculty of Health Sciences, UiT - the Arctic University of Norway, NO-9037 Tromsø, Norway – Norway

[2] Department of Medicinal Chemistry, Institute of Pharmacology, Polish Academy of Sciences, 12 Smetna Street, Kraków 31-343, Poland – Poland

-aminobutyric acid (GABA) is the main inhibitory neurotransmitter in the central nervous system (CNS), and dysregulation of the GABAergic system is related to brain disorders[1]. The GABABreceptor is a heterodimeric class C G-protein coupled receptor (GPCR) consisting of two subunits (GABAB1and GABAB2) [2]. GPCRs are targets for more than 1/3 of marketed drugs. Most of these drugs bind to the orthosteric site, but due to the structural conservation of the orthosteric binding site among the GPCRs they may lack selectivity. Allosteric modulators (AMs) have higher selectivity than regular orthosteric drugs and hence may trigger fewer side effects. For GABAB receptor, the allosteric binding pocket is located within the 7TM bundle of GABAB2 [3,4].

No experimental structures of the transmembrane domain are available. Thus, with the technique of homology modelling we have generated several hundred models of GABAB2using templates from different GPCR families; the closest having _˜20% sequence identity. The modelling was guided by the capacity of the models to enrich clustered known AMs. The models went into one round of Induced-Fit Docking (IFD) in order to increase their ligand-specificity. The evaluation of the selected models indicated that they complied well with available mutagenesis data and important residues were identified [5]. The GABAB2 models were used as tools in a structure-based virtual ligand screening for new allosteric GABABmodulators.

Acknowledgements

The study was partially supported by the Polish-Norwegian Research Programme operated by the National Centre for Research and Development under the Norwegian Financial Mechanism 2009-2014 in the frame of the Project PLATFORMex (Pol-Nor/198887/73/2013).

Thibaud Freyd is a fellow of the National graduate school in structural biology (BioStruct).

References

1. Lehmann K, Steinecke A, Bolz J. Neural Plast. 2012;2012: 892784.

*Speaker

†Corresponding author: ingebrigt.sylte@uit.no

2. Brown KM, Roy KK, Hockerman GH, Doerksen RJ, Colby DA. J Med Chem. 2015;58: 6336–6347.

3. Binet V, Brajon C, Corre LL, Acher F, Pin J-P, Prézeau L. J Biol Chem. 2004;279: 29085–29091.

4. Dupuis DS, Relkovic D, Lhuillier L, Mosbacher J, Kaupmann K. Mol Pharmacol. 2006;70: 2027–2036.

5. Freyd T, Warszycki D, Mordalski S, Bojarski AJ, Sylte I, Gabrielsen M. PLOS ONE. 2017;12: e0173889.

Keywords: drug, virtual screening, homology modelling, gabab

Amino acid pair interactions in membrane proteins investigated with the statistical potential formalism

Mame Ndew Mbaye * [1]

[1] Université Libre de Bruxelles [Bruxelles] (ULB) – Avenue Franklin Roosevelt 50 - 1050 Bruxelles, Belgium

Amino acid pair interactions in membrane proteins investigated with the statistical potential formalism

M.N. Mbaye, Q. Hou, F. Pucci⋆, M. Rooman⋆

Department of BioModeling, BioInformatics and BioProcesses, Universite Libre de Bruxelles, CP 165/61, Roosevelt Ave. 50, 1050 Brussels (⋆) Co-last authors

Abstract

Using the statistical potential formalism, we developed new energy functions describing amino acid pair interactions in membrane proteins, in which we made the distinction between transmembrane and intra/extra-cellular regions. The comparison of these potentials with potentials derived from globular proteins led us to objectively identify and interpret the key interactions in the different protein environments.

Introduction

Membrane proteins are a very important class of proteins, whose structure and composi- tion substantially differ from globular proteins due to their incorporation into biological membranes, mainly composed of hydrophobic lipid molecules. They play important roles in cellular function by transferring molecules, ions and different types of signals from the cell exterior to the interior and vice versa, as well as in the localization and organization of the cell. They constitute about 30% of the entire human proteome [1]. They are the focus of a lot of pharmaceutical research, as they correspond to about 60% of the current drug targets [2].

The folding, stability and activity of membrane proteins is reached only within the lipid bilayer, which complicates getting their experimental X-ray structures. Generally, their large size makes also difficult to obtain them by nuclear magnetic resonance spectroscopy. These are the reasons why transmembrane protein structures only represent about 2% of the available structures deposited in the Protein Data Bank (PDB) [3]. The analysis and modeling of the 3-dimensional (3D) structure of membrane proteins are thus key objectives, for example in view

*Speaker

of rationally guiding protein design and engineering experiments. In spite of their importance, membrane proteins have been much less studied than globular proteins.

Membrane and globular protein datasets

To set up our membrane protein dataset, we used the OPM database [4], which contains experimental structures of membrane proteins. From these, we selected the structures obtained by X-ray crystallography with a resolution of 2.5 A at most. Our dataset D is a subset of these structures where we imposed a threshold on the pairwise sequence identity of 30% with the help of the protein sequence culling server PISCES [5]. It con- tains 170 structures, which were then divided into their transmembrane, extracellular and cytoplasmic regions using the OPM annotations. We got in this way the dataset DTM, which contains the transmembrane protein parts, and DEC that mixes the extracellular and cytoplasmic regions.

For comparison, we also considered the DG dataset set up in [6], which contains 3,823 X-ray structures of globular proteins, with a resolution of 2.5 A at most and a pairwise sequence identity lower than 20 %.

The amino acid frequencies differ in these datasets for some amino acids. The clearest difference is observed for the aliphatic residues Val, Ile and Leu: in DEC, they are most often in the core and rarely at the surface (as in globular protein) whereas in DTM, they are almost evenly distributed between core and surface, with a slight preference for partially buried regions. Note that there are much more aliphatic residues in transmembrane than in extra/intra-cellular regions.

In contrast, charged amino acids are much more frequent in DEC than in DTM; in both sets they are consistently more often at the surface than in the core. One should here distinguish between the exterior and cytoplasmic regions, as positively charged amino acids are known to be more frequent at the cytoplasmic side, near the interface with the membrane where they interact with lipid head groups [7].

Statistical residue-residue potentials

To analyze objectively the residue-residue interactions that contribute to the stability of transmembrane or intra/extra-cellular regions, we used the statistical potential formalism. Statistical potentials are coarse-grained energy functions derived from frequencies of ob- servation of sequence-structure associations in a structure dataset. Here we considered distance-dependent residue-residue potentials defined as in [8, 9]:

$$W(s1,s2,d) = -kBT \ln [F(s1,s2,d) \ (1)/ \ F \ (s1, s2)F \ (d)]$$

where s1 and s2 are amino acid types and d is their spatial distance computed between the average side chain geometric centers. The distances are discretized into distance bins and the residue pairs are restricted to those that are separated by at least 8 residues along the sequence [9]. T is the absolute temperature and kB the Boltzmann constant. The relative frequencies F(s1,s2,d), F(s1,s2) and F(d) are computed in a specific dataset. Here we considered the three datasets, DTM, DEC and DG, and computed from them the three potentials WTM, WEC and WG.

The Glu-Lys pair potential, representing salt bridge interactions, is almost identical for extra/intra-cellular regions and globular proteins and much more favorable for transmembrane regions. This means that Glu and Lys have the clear tendency to form salt bridges when they are in the membrane, whereas outside the membrane they only have this tendency when they are buried in the core; at the surface they make favorable interactions with water molecules. Note that this trend is independent of the frequency of Glu and Lys, which is lower inside the membrane.

For the aliphatic residue pairs Val and Leu, we again find that the potentials for intra/extra-cellular regions and globular proteins almost coincide, but the transmembrane potential is here less favorable. This means that these residues, which are hydrophobic, make stabilizing contacts only when the protein is surrounded by water, not by a hydrophobic medium. The same is true for pairs of aromatics Phe residues, which are also hydrophobic. The difference here is that Phe-Phe interactions are more favorable in globular proteins than in intra/extra-cellular regions.

For some residue pairs, the difference between the potentials W T M , W EC and W G is small. This is the case for example for the Phe-Glu pair, which represents anion-π interactions. This potential is, however, somewhat more favorable at short distance in transmembrane regions.

Conclusion

The transmembrane and intra/extra-cellular residue-residue potentials that we developed allow the identification of the amino acid interactions that stabilize transmembrane and extra/intra-cellular regions, and their comparison with interactions in globular proteins. In a next step, we will use them to predict whether a protein region is situated inside or outside the membrane, and the effect of mutations on membrane protein stability.

References

Fagerberg L, Jonasson K, von Heijne G, Uhln M, Berglund L. Prediction of the human membrane proteome. Proteomics. 2010; 10:1141-9.

Bakheet TM, Doig AJ. Properties and identification of human protein drug targets. Bioinformatics. 2009; 25:451-7.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235-42.

Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. 2012; 40:D370-6.

Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003; 19:1589-91.

Pucci F, Bourgeas R, Rooman M. Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. Sci Rep. 2016; 6:23257.

De Marothy MT, Elofsson A. Marginally hydrophobic transmembrane α-helices shaping membrane protein folding. Protein Sci. 2015; 24:1057-74.

Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol. 1990; 213:859-83.

Kocher JP, Rooman MJ, Wodak SJ. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J Mol Biol. 1994; 235:1598-613.

Keywords: Amino acid pair interactions, membrane proteins, statistical potential

Solubility-dependent statistical potentials to probe into the (in)solubilizing effect of residue-residue interactions in proteins

Hou Qingzhen *† [1], Raphaël Bourgeas , Fabrizio Pucci , Marianne Rooman

[1] Department of BioModeling BioInformatics BioProcesses, Université Libre de Bruxelles, Brussels, Belgium – Belgium

(a full version including formulas and figures can be seen in the pdf uploaded)

Abstract

Solubility is a basic biophysical property of globular proteins, which is often crucial for their correct functioning. Here we investigated the protein solubility properties in the framework of the statistical potential formalism. We derived solubility-dependent distance potentials, which led to an objective determination of the solubility dependence of specific amino acid interactions. We also computed the protein folding free energy with these potentials to investigate whether they can be used for solubility prediction.

Introduction

The understanding of the mechanisms that modulate protein solubility and aggregation of globular proteins is quite complicated due to their dependence on many intrinsic and extrinsic factors. They are intimately connected with the stability of their three-dimensional (3D) structure but also vary according to the environmental conditions. Unraveling these complex relationships is a crucial objective for many academic and biotechnological applications. Indeed, insolubility is frequently a bottleneck in high-throughput structural genomic studies and in applications requiring high-concentration production of recombinant proteins, such as monoclonal antibody solutions for pharmaceutical applications [1, 2]. Despite some important advances, the precise identification of the amino acid interactions and structural characteristics that ensure solubility or insolubility and their biophysical interpretation remain elusive.

Protein structure and solubility dataset

We constructed a dataset of globular proteins with experimentally characterized structure and solubility. We used for this purpose the eSOL database [3], containing proteins of which the solubility S was measured as the ratio of the supernatant protein fraction obtained after centrifugation of the translation mixture, and the total uncentrifuged protein fraction. We mapped the eSOL entries onto the corresponding 3D structures in the Protein Data Bank [4]. The protein-culling server PISCES [5] was then used to remove proteins with similar sequences ($\geq$ 25% identity) and low-resolution structures ($>$ 2.5 Å).

The Dtot set so obtained contains 412 proteins. We divided it into two subsets with an equal number of proteins: Dsol which contains all structures with solubility S $\geq$ 64%, and D$insol$ with S $<$ 64%.

*Speaker

†Corresponding author: qingzhen.hou@ulb.ac.be

Solubility-dependent distance potentials
We used the statistical potential formalism to describe the interaction strength between two amino acid of type s and s 0 as a function of their distance d, computed between the average geometric center of their side chains. The common distance potentials are derived from the frequency of observation F of associations between s, s 0 and d in a structure dataset – here D tot –, using the inverse Boltzmann law [6]: Formula(1)
where k B is the Boltzmann constant and T the absolute temperature.
In addition, we defined two novel potentials aimed at representing the solubility properties of the proteins from which they are derived. They were obtained from the D sol and D insol datasets, as well as the full D tot set, by generalizing the approach of [7]: Formula(2)
These solubility-dependent potentials were utilized to quantify the contribution of amino acid pair interactions to protein solubility.

(In)solubilizing effect of residue-residue interactions
The folding free energy profiles as a function of the inter-distance d, obtained with the three potentials defined in Eqs (1-2), were computed and analyzed for all 210 possible residue-residue pairs. Their comparison led to the identification of the residue pairs for which the profiles differ significantly, and thus of the interactions that contribute more strongly than the others to the increase or decrease of protein solubility.
We found that the soluble and insoluble folding free energy profiles obtained with W sol and W insol differ for a large number of residue pairs, with the W tot profiles between the two. We grouped and analyzed together the residue pairs that share similar biophysical characteristics, in order to rationalize their contribution to protein (in)solubility and interpret the underlying physical principles. The seven group potentials are shown in Fig. 1.
The analysis of this figure indicates that the interactions that are more favorable in insoluble proteins than in soluble proteins are: (1) aromatic-aromatic interactions (Fig. 1.A), (2) His-aromatic interactions (Fig. 1.B), (3) cation-π interactions involving arginine (Fig. 1.C), (4) amino-π interactions (Fig. 1.D), and (5) anion-π interactions (Fig. 1.E).
These findings highlight the role of charge delocalization in the insolubility or aggregation properties of globular proteins. Indeed, all the interactions that involve residues with delocalized π-electrons on their side chain disfavor solubility. This is the case of the aromatic residues Phe, Tyr and Trp, of the aromatic and sometimes positively charged residue His, of the positively charged Arg, of Gln and Asn that possess a side chain amide group, and of the negatively charged residues Asp and Glu.
In contrast, the residue pairs for which the potential derived from soluble proteins is significantly more favorable than the potential derived from insoluble proteins are: (1) aliphatic-aliphatic interactions (Fig. 1.F), and (2) Lys-containing salt bridges (Fig. 1.G). Thus, the side chain interactions that promote protein solubility do not involve delocalized π-electrons.
Solubility from solubility-dependent potentials
To test how the energies computed with the solubility-dependent statistical potentials correlate with solubility, we computed three folding free energy values for each protein from the D tot set of sequence S and conformation C: Formula (3).where s i and s j are two amino acid types at positions i and j along the sequence, N the sequence length and α =tot, sol or insol. To avoid any overfitting, the folding free energies were computed using a leave-one-out cross validation strategy. We also computed the soluble and insoluble folding free energy difference: Formula (4).

As shown in Table 1, the Pearson correlation coefficient of the experimental solubility with the folding free energy difference is quite good (0.4). It is better than the correlation with the usual folding free energy W tot (S, C), as well as with commonly used sequence features, namely the protein length, the isoelectric point and the aliphatic index.
Conclusion

Our new solubility-dependent mean force potentials were used to clarify the relation between the amino acid interactions and the solubility propensities, and led to the interesting result that interactions involving delocalized electrons promote insolubility, whereas others ensure solubility. Moreover, we found that the folding free energy of proteins computed with these potentials correlate well with experimental solubility – better than commonly used features –, which lets foresee that predictors based on these potentials will outperform existing solubility predictors.
References

F. Baneyx, M. Mujacic, Recombinant protein folding and misfolding in Escherichia coli, Nat Biotechnol. 22 (2004) 1399-408.

S.S. Mohan, P.A. Kumar, Solubilization and refolding of bacterial inclusion body proteins, Journal of Bioscience and Bioengineering 99 (2005) 303-310.

T. Niwa, B.W. Ying, K. Saito, W.Z. Jin, S. Takada, T. Ueda, H. Taguchi, Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins 106 (2009) 4201-4206.

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The Protein Data Bank. Nucleic Acids Res. 28 (2000) 235-42.

G. Wang, RL Jr. Dunbrack. PISCES: a protein sequence culling server. Bioinformatics. 19 (2003) 1589-91.

Kocher JP, Rooman MJ, Wodak SJ. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J Mol Biol. 1994; 235:1598-613.

B. Folch, Y. Dehouck, M. Rooman, Thermo- and mesostabilizing protein interactions identified by temperature-dependent statistical potentials, Biophys J. 98 (2010) 667-77.

Keywords: protein solubility, statistical potential, amino, acid interactions, solubility, dependent potentials

Identification of additional Ser/Thr protein kinases in the genome of Streptococcus thermophilus using structural homology detection.

Samantha Samson *† 1, Véronique Martin 1, Lucia Haller 2, Véronique Monnet 3, Gwenaëlle André-Leroux 1

[1] Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE) – Institut national de la recherche agronomique (INRA) – INRA - Unité MaIAGE Bât. 233 et 210 Domaine de Vilvert, 78352, JOUY-EN-JOSAS CEDEX, France
[2] Plate-forme PAPPSO, Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay (PAPPSO) – Institut national de la recherche agronomique (INRA) – INRA - Domaine de Vilvert, 78352, JOUY-EN-JOSAS CEDEX, France
[3] MICrobiologie de l'ALImentation au Service de la Santé humaine (MICALIS) – AgroParisTech, Institut national de la recherche agronomique (INRA) – INRA - Domaine de Vilvert, 78352, JOUY-EN-JOSAS CEDEX, France

A protein function is mainly determined by its 3-dimensional structure, which is much more preserved than the amino acid sequence. Given that the folding of a protein correlates to its sequence, the structural prediction of a protein of unknown function paired with multiple-sequence-alignment and amino-acid frequency could lead to the identification of its function. Reversely, a protein expected to have a specific function can been retrieved through its associated prediction of 3D structure. These properties can be scaled up to an entire genome. This is the issue we address here.
In this project, we aim to identify structural homologues of the Ser/Thr protein kinase PknB. Indeed, the phospho-proteomes of the lactic acid bacterium *Streptococcus thermophilus* LMD-9 native and PknB, that were performed and analyzed within the PAPPSO* facility, evidenced only a slight decrease of 10% of phosphorylated proteins (1). This redundancy clearly drives the hypothesis of additional kinases, present in the genome, despite the solely PknB Ser/Thr kinase annotated in the genome. Using MetaFoldScan, a tool developed in MaIAGE** to screen whole genomes, and targeting the canonical catalytic kinase domain of PknB from *Mycobacterium tuberculosis* (2) as a structural bait, we screened the entire *S. thermophilus LMD-9* genome to identify structural homologues of PknB kinase domain.

The protocol follows the steps :

1. Splitting of *S. thermophilus* genome, multiple sequence alignment using the uniprot20_2016_02 database, and prediction of secondary elements.

2. Hidden Markov Model (HMM) profiling (3) for each sequence and for *M. tuberculosis* PknB

[*] Speaker

[†] Corresponding author: samantha.samson@inra.fr

target.

3. HMM profile-profile comparison with scoring and ranking of the results according to the probability of structural homology.

Eventually, among the 1673 protein sequences identified from the genome, we spotted 3 structural hits as putative kinases, including one annotated as hypothetical protein. The three showed the higher probability of being structural homologues of PknB kinase domain even with less than 12% pairwise sequence identity (percentage of residues identical between two proteins), and thus they were subsequently analyzed using Modeller (4) and PyMOL (5).

We now aim to characterize *in vitro* the phenotype and phosphoproteome corresponding to the depletion of this most relevant hit.

Bibliography:

1. Haller L, Henry C, Blein-Nicolas M, Zivy M, Canette A, Verbrugghe M, Mézange C, Boulay M, Monnet V. The Hanks-type kinase PknB, targeting the divisome, is not the single player in the Streptococcus thermophilus protein phosphorylation process (To be published).

2. Ortiz-Lombardía M, Pompeo F, Boitel B, Alzari PM. Crystal structure of the catalytic domain of the PknB serine/threonine kinase from Mycobacterium uberculosis. J. Biol. Chem. 2003. 278 (15): 13094-100.

3. S´oding J, Biegert A, Lupas N. The HHpred interactive server for protein detection and structure prediction. Nucleic Acids Res. 2005 Jul 1; 33(Web Server issue): W244–W248.

4. Sali A. & Blundell. T.L. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol 1993. 234, 779-815.

5. DeLano, W.L. The PyMOL Molecular Graphics System 2002. http://www.pymol.org.

* PAPPSO : Plateforme d'Analyse Protéomique de Paris Sud-Ouest : http://pappso.inra.fr.
** MaIAGE : "Mathématique et Informatique Appliquées du Génome à l'Environnement", Unité de recherche - INRA

Keywords: Protein structure prediction, Streptococcus thermophilus LMD 9, Ser/Thr protein kinase

Involvement of 3D motifs in RNA secondary structure prediction

Phuong Chu * [1], Audrey Legendre [1], Eric Angel [1], Fariza Tahi† [1]

[1] IBISC, Univ Evry, Université Paris Saclay (IBISC) – Université d'Evry-Val d'Essonne : EA4526 - 23 – Bd de France 91034 – EVRY, France

RNAs are molecules performing a broad range of functions in cells, which depend on the folding properties of these molecules. Therefore, RNA structure prediction is an important field in bioinformatics. RNA structures are typically described at two levels of organization: the secondary and tertiary structures. The secondary structure, which contains canonical base pairs (i.e. Watson-Crick A-U, G-C and Wobble G-U base pairs), determines the backbone of the molecular structure. Studying secondary structures and base pairing properties can reveal fundamental insights into the functional mechanisms of RNAs such as frame shift elements [1] and riboswitches [2]. The tertiary structure indicates the three-dimensional structure positions of each atom in the molecule. Both canonical and non-canonical interactions (classified by Leontis and Westhof [3]) are included in the tertiary structures. This information is essential to gain in-depth information about the functions. Most of strategies and tools have been proposed to obtain RNA secondary structures because predicting secondary structures is simpler than predicting tertiary structures. However, sometimes they still fail to return genuine structures. In order to obtain the structures that are closer to the real structure, we consider using information of common 3D motifs in the prediction of RNA secondary structure. At the moment, there is only one tool, called RNAMoIp [4], which performs the modification of RNA secondary structures in order to include the motifs. Therefore, this study is a quick assesment on the role of 3D motifs on the prediction of RNA secondary structures. We investigate the presence of 3D motifs inside the RNA sequence and how often they appear; at the same time, we consider the potential of using this information in RNA secondary structure prediction.

Our study is based on a dataset of RNAs without pseudoknots because most of the popular RNA secondary structure prediction tools only give back results without pseudoknots. These RNAs sequences were gathered from the RNA STRAND database v2.0 [5]. This dataset includes 145 sequences whose lengths range from 10 to 97 nucleotides.

RNA motifs are identified and described as recurrent short sequences in functional RNAs. In the secondary structures, a motif can be represented in 3 basic shapes containing no base pairs: hairpins, interior loops and k-way junctions. A 3D motif is formed by non-canonical interactions. These interactions complete the folding of the RNAs, allowing them to hold their functions. The information of motifs can be extracted from the RNA three-dimensional structures. We used the motifs database provided together with RNAMoIP program [4], which includes 4655 non-redundant RNA motifs. The database follows the format in RNA3DMotif software of Djelloul

*Speaker

†Corresponding author: fariza.tahi@univ-evry.fr

and Denis [6]. The format consists in splitting a motif into components. A component is a short sequence that is part of the motif. Each component is set to have a distance of at least 3 nucleotides from each other. Then a hairpin has 1 component, an interior loop has 2 components and a k-ways junction has k components.

We used RNAsubopt from ViennaRNA Package [7] and Biokop [8] to generate 10 secondary structures from the RNA sequences (one optimal solution and 9 sub-optimal solutions). RNAsubopt ranks different solutions based on minimum free energy (MFE) model. Biokop uses a multicriteria approach to rank the solutions.

Both the RNA sequences and their secondary structures are treated as our input data. Our tool analyses the role of RNA motifs by two steps:
(1) Apply a classical pattern-matching algorithm to find all possible motifs that can be included inside the RNA sequence.
(2) Apply a second pattern matching on the predicted secondary structures to find which secondary structures contain the identified motifs.

We first study the matching of 3D motifs in referenced sequences. We observed that 51.72% of referenced structures did not contain any motif, 44.83 % of referenced structures contain 1 motif and 3.45 % of referenced structures contain 2 motifs. There is not any referenced structure that contains more than 2 motifs. We also observed that all of the found motifs are hairpin motifs.

After applying our proposed method, 10 first RNA secondary structures of each RNA sequence are assigned into different groups based on their numbers of contained motifs. In relatively short sequences (less than 50 nucleotides) , both RNAsubopt (Figure 1) and Biokop (Figure 2) work well on returning the best structure (i.e the structure with the highest accuracy is also the first returned structure or optimal solution). However, when the sequences are longer (more than 50 nucleotides), best secondary structures are hidden inside the set of sub-optimal solutions. Overall, the percentage when the optimal solutions returned by RNAsubopt have the highest accuracy is 63.44% while this percentage in Biokop is 36.55%.

In general, those solutions, which are returned by both programs, do not have more than 2 motifs, and those found motifs are hairpin motifs – which corresponds to our first observation on the referenced structures. The only exception is the solution sub-optimal#8 of PDB_00547 (returned by RNAsubopt), which contains 3 motifs. In our dataset, the sequence PDB_00547 is the longest sequence with the length 97 nucleotides; so the probabilities to include more motifs are higher than other sequences. The best solutions (secondary structures with the highest accuracy), in most cases, contain motifs: 75.17% in RNAsubopt and 81.38 % in Biokop. When we categorize all obtimal and suboptimal solutions into new smaller groups (which contains different numbers of motifs), we can clearly see the differences in accuracy among those groups. We observe that structures contained motifs have higher accuracy than those without.

The results obtained by our method show that taking into account RNA motifs would significantly improve the accuracy of RNA secondary structure prediction. At the same time, it is important to update the motif databases in order to maximize the benefit from RNA motifs information.

References

[1] Belkaert,M., et al (2003) Towards a computational model for -1 eukaryotic frameshifting sites. Bioinformatics. 2003 Feb 12;19(3):327-35.

[2] Vitreschak,A.G., Rodionov,D.A., Mironov,A.A., Gelfand,M.S. (2004) Riboswitches: the oldest mechanism for the regulation of gene expression?. Trends Genet. 2004 Jan;20(1):44-50.

[3] Leontis,N. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. RNA, 7, 499-512.

[4] Reinharz,V., Major,F., and Waldispuhl,J. (2012) Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure. Bioinformatics. 2012 Jun 15;28(12):i207-14.

[5] Andronescu,M., Bereg,V., Hoos,H.H., and Condon,A., (2008) RNA STRAND: The RNA Secondary Structure And Statistical Analysis Database. BMC Bioinformatics. 2008;9(1):340.

[6] Djelloul,M. and Denise,A. (2008) Automated motif extraction and classification in RNA tertiary structures. RNA. 2008 Dec;14(12):2489-97.

[7] Lorenz,R., et al (2011) "ViennaRNA Package 2.0", Algorithms for Molecular Biology: 6:26.

[8] Legendre,A., Angel,E., and Tahi,F. (2018). Bi-objective integer programming for RNA secondary structure prediction with pseudoknots. BMC bioinformatics, 19(1), 13.

Keywords: RNA, RNA structure, RNA Secondary structure prediction, 3D Motifs

Evaluation de l'assemblage de données Rna-seq et de la construction des bases de données protéogénomiques d' organismes non modèles

Yannick Cogne *† 1, Duarte Gouveia 1, Arnaud Chaumot 2, Olivier Geffard 2, Olivier Pible 1, Jean Armengaud‡ 1, Christine Almunia§ 1

[1] Innovative technologies for Detection and Diagnostics (Li2D) – CEA Marcoule – BP17171, F-30207 BAGNOLS-SUR-CEZE, France

[2] Milieux aquatiques, écologie et pollutions (UR MALY) – CEMAGREF – 5 rue de la Doua, CS70077, 69626 Villeurbanne Cedex, France, France

L'utilisation des nouvelles méthodologies omiques en écologie et écotoxicologie booste les avancées scientifiques dans ces domaines d'actualités. L'évaluation sur le vivant et les écosystèmes des impacts des polluants tels que métaux lourds, nanoparticules, agents chimiques cancérigènes ou perturbateurs endocriniens peut être effectué à l'aide d'organismes sentinelles. L'encagement d'individus représentatifs et bien calibrés sur un site pollué et l'analyse moléculaire la plus fine possible permet de renseigner le niveau de pollution global des toxiques biodisponibles sur la Vie. Par exemple, des paramètres physiologiques tels que des retards de mue ou de développement de leurs appareils reproducteurs peuvent être quantifiés par le biais de marqueurs protéiques spécifiques. Cette vision intégrative est indispensable pour une meilleure gestion de notre environnement malheureusement trop fortement impacté par l'Homme.

Nos équipes de recherche ont proposé le modèle Gammare, un petit amphipode aquatique, et des stratégies de recherche de biomarqueurs protéiques des plus raffinées afin de rationaliser les analyses d'impact de polluants présents dans les rivières. Sur l'espèce *Gammarus fossarum*, des analyses RNAseq et protéomique à très haut-débit ont permis d'établir une base de données protéogénomiques très détaillée (Trapp et al. 2014 ; Trapp et al. 2016). Une analyse fine des réponses moléculaires de l'organisme sentinelle soumis à différents perturbateurs endocriniens a été obtenue par l'analyse protéomique d'individus soumis à différents stress (Trapp et al. 2015). Les biomarqueurs sélectionnés sur cette base de nouvelles connaissances ont été validés par leur dosage systématique de grandes cohortes d'individu possible par une méthodologie de protéomique ciblée utilisant le mode " selected reaction monitoring " de la spectrométrie de masse en tandem (Charnot et al., 2017 ; Gouveia et al., 2017a). Désormais, cette méthode de quantification a été appliquée sur des échantillons *in natura* prélevés sur de multiples cours d'eau de rivières du Sud-Est de la France. Les résultats de cette méthodologie multi-omique intégrative permettent d'établir un indicateur de santé des rivières (Gouveia et al., 2017b).

*Speaker

†Corresponding author: yannick.cogne@hotmail.fr

‡Corresponding author: jean.armengaud@cea.fr

§Corresponding author: christine.almunia@cea.fr

Dans ce type de campagne de protéogénomique qui peut s'appliquer à tout type d'organismes vivants (Armengaud et al., 2014), l'interprétation des millions de spectres MS/MS générés par spectrométrie de masse en tandem requiert une base de données listant les séquences de toutes les protéines théoriquement présentes. Ces bases de données sont issues des génomes annotées pour les organismes modèles (tels que *Homo Sapiens, Escherichia coli, Saccharomyces cerevisiae*) mais ne sont que partielles lorsqu'il s'agit d'autres organismes peu caractérisés génomiquement ou de lignée cellulaire anormale (cancéreuse), voire inexistantes pour la plupart des animaux utilisés en écotoxicologie. D'autre part, la très forte variabilité génomique des animaux prélevés *in natura* rend difficile la constitution de base de données de qualité et donc cette interprétation. Afin d'obtenir cette information cruciale, des données de séquençage transcriptomique à haut débit (RNAseq) peuvent être assemblées puis traduites. La base de données protéomiques ainsi créée comprend les bonnes séquences des polypeptides présents mais aussi certaines séquences aberrantes. Les logiciels d'interprétation des données expérimentales de protéomique permettent de séparer le bon grain de l'ivraie, *i.e.* différencier les polypeptides effectivement présents dans les échantillons de ceux qui ne sont pas présents ou qui correspondent à des séquences aberrantes. La méthodologie de protéogénomique appliquée à des organismes non-modèles tels que *Gammarus fossarum* permet donc de tirer le meilleur parti des données RNAseq et protéomique nouvelle génération. Ainsi, des biomarqueurs spécifiques de l'impact de polluants peuvent être mis en évidence rapidement. Toutefois, *in natura*, il existe de nombreuses espèces de Gammares, et au sein de l'espèce *G. fossarum* de nombreux sous-groupes génétiquement divers peuvent être définis. Il est donc important de prendre en compte cette diversité génomique et d'évaluer la transversalité des biomarqueurs d'intérêt.

Afin de développer nos connaissances sur la diversité moléculaire des biomarqueurs d'intérêt en écotoxicologie, nous nous proposons d'étudier des représentants de différents groupes taxonomiques de Gammares et différentes populations de la même espèce. Des jeux de données de protéomique shotgun de dernière génération et de RNA-seq exhaustif ont été produits sur des individus génotypés. Huit individus mâles et huit individus femelles de chacun des groupes ont été testés en protéomique. Un total de 16 transcriptomes différents sur 16 individus génotypés représentatifs des différents groupes ont été obtenus. Le premier objectif est d'assembler ces 16 transcriptomes alors qu'aucun génome de référence suffisamment proche phylogénétiquement n'existe. Le deuxième objectif est de pouvoir interpréter les données de protéomique d'une cohorte très importante d'invidus, et ce avec suffisamment de sensibilité pour pouvoir détecter par la suite les différences entre chaque groupe. Alors que pour l'utilisation des données Rna-seq dans le cadre d'organisme modèle des outils ont été développé (Madar et al., 2017, Wen et al. 2016), aucune solution intégrale n'est proposée pour l'étude protéogénomique d'organisme non modèle. Pour cela, les protocoles d'assemblage de transcriptome actuellement recommandé dans le cadre d'assemblage de RNAseq d'organismes modèles ont été réexaminés à l'aune de la problématique des organismes non-modèles et de leur exploitation par protéogénomique. L'impact des pré-traitements des données RNAseq obtenus par la technologie Illumina sur la qualité de l'assemblage a été évalué par les outils classiques tels que i) mesures de longueurs de séquences et ii) évaluation de la présence des séquences similaires avec les plus proches taxons. Nos résultats démontrent que ces paramètres ne permettent pas de différencier facilement l'impact des différents pré-traitements possibles de données RNAseq. *A contrario*, l'interprétation protéogénomique de données protéomiques obtenus sur des individus *Gammarus fossarum groupe A* femelles génotypées permet de rapidement sonder la qualité des assemblages testés. L'impact des différents traitements de données RNAseq sera détaillé. D'autre part, la création de la base de données protéogénomique à partir de l'assemblage RNAseq peut être également optimisée. Nous avons exploré différentes options de paramètres de recherche des séquences codantes à partir des transcrits de RNAseq en se basant sur l'outil de recherche d'ORF Transdecoder. L'optimisation de ces paramètres permet de réduire considérablement l'espace de recherche tout en maximisant le taux d'identification des spectres MS/MS et en diminuant fortement le temps de requête et la puissance de calcul nécessaire pour de grands jeux de données.

L'ensemble de ces résultats obtenus sur un organisme non-modèle de référence, *G. fossarum*, nous permet aujourd'hui de présenter un protocole optimisé d'assemblage de données RNAseq et de construction de base de données pour la protéogénomique d'organismes non-modèles.

References :

Gouveia D et al. (2017) Ecotoxico-Proteomics for Aquatic Environmental Monitoring: First in Situ Application of a New Proteomics-Based Multibiomarker Assay Using Caged Amphipods. Environ Sci Technol. 51(22):13417-13426.

Gouveia D et al. (2017) Assessing the relevance of a multiplexed methodology for proteomic biomarker measurement in the invertebrate species Gammarus fossarum: A physiological and ecotoxicological study. Aquat Toxicol. 190:199-209.

Charnot A et al. (2017) Multiplexed assay for protein quantitation in the invertebrate Gammarus fossarum by liquid chromatography coupled to tandem mass spectrometry. Anal Bioanal Chem. 409(16):3969-3991.

Trapp J et al. (2016) High-throughput proteome dynamics for discovery of key proteins in sentinel species: Unsuspected vitellogenins diversity in the crustacean Gammarus fossarum. J Proteomics. 146:207-14.

Trapp J et al. (2015) Proteomic investigation of male Gammarus fossarum, a freshwater crustacean, in response to endocrine disruptors. J Proteome Res. 14:292-303.

Trapp J et al. (2014) Proteogenomics of Gammarus fossarum to document the reproductive system of amphipods. Mol Cell Proteomics. 13:3612-25.

Armengaud J et al. (2014) Non-model organisms, a species endangered by proteogenomics. J Proteomics. 105:5-18.

Wen, Bo, et al. (2016) "PGA: an R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq." BMC bioinformatics 17.1: 244.
Madar, et al. (2017) "Comprehensive and sensitive proteogenomics data analysis strategy based on complementary multi-stage database search." International Journal of Mass Spectrometry

Keywords: Proteomics, Transcriptomics, Benchmarking, Rna, seq, Proteogenomics

Analyzing RNA folding landscapes using non-redundant sampling

Juraj Michalik [*] [1], Yann Ponty[†] [1], Christelle Rovetta [1], Hélène Touzet [2]

[1] Laboratoire dínformatique de l´cole polytechnique [Palaiseau] (LIX) – Centre National de la Recherche Scientifique : UMR7161, L'Institut National de Recherche en Informatique et e n Automatique (INRIA) – 1 rue Honoré d'Estienne d'Orves Bâtiment Alan Turing Campus de l'École Polytechnique 91120 Palaiseau, France

[2] Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL) – CNRS : UMR9189 – Université de Lille, campus cité scientifique, Bâtiment M3 extension, Avenue Carl Gauss, 59655 Villeneuve d'Ascq CEDEX, France

Kinetics is key to understand many facets of structural non-coding RNAs, including co-transcriptional folding and riboswitches. Exact out-of-equilibrium studies typically induce extreme computational demands, leading state-of-the-art methods to rely on approximated kinetics landscapes. Those are obtained using sampling strategies that strive to generate the key landmarks of the landscape topology. However, such methods are impeded by a large level of redundancy within sampled sets. Such a redundancy is uninformative, and obfuscates important intermediate states, leading to an incomplete vision of RNA dynamics.

In a recent work [1], we introduce RNANR, a new set of algorithms for the exploration of RNA kinetic landscapes at the secondary structure level. These algorithms are accessible at: https://project.inria.fr/rnalands/rnanr

Approach and state of the art. RNANR considers locally optimal structures, or local optima, a reduced set of RNA conformations, in order to focus its sampling on the basins of the kinetic landscape. Accordingly, our definition of locally optimal structures coincides with that of saturated structures, for which no base pair can be added without creating i) a crossing base-pair, aka pseudoknot, excluded from RNANR due to complexity reasons; and ii) base triplets, so each base can appear in at most one base pair. We then built on an exhaustive enumeration algorithm contributed by Saffarian et al [2], and contributed a statistical sampling algorithm within local minima using dynamic programming. This contrasts with the majority of current approaches that

[*] Speaker

[†] Corresponding author: yann.ponty@lix.polytechnique.fr

either enumerate all suboptimal structures (structures with higher energy than global optimum) within some energy range (RNASLOpt [3], RNAsubopt [4]) or use some variation of Boltzmann-Gibbs sampling from the set of all secondary structures (RNAlocopt [5]). RNAsubopt also served as a base for simulated annealing-inspired algorithm used in RNAlocmin [6]. An approach similar to the the exhaustive enumeration algorithm used in RNANR was also independently designed by Waldisp´uhl et al [7].

Method. Along with an exhaustive enumeration, RNANR implements a novel non-redundant stochastic sampling, and offers a rich array of parameters to imprint expert knowledge in refining the definition of relevant locally optimal structures. Sampling probabilities for any specific structure are computed according to the Boltzmann distribution, giving more importance to lower free-energy states, while still allowing the generation of higher free-energy structures. Our energy model is the nearest neighbor model contributed by the Turner group [8]. Moreover, we forbid any redundancy within our sampling, thanks to the design of a dedicated tree-like data-structure. This structure remembers not only all local optima that were already sampled, but also the chain of intermediate steps leading to them, consisting in the choice of a given sub-motif for any corresponding interval. Should a decision potentially lead to previously sampled structure, it sees its probability modified using the values stored in data-structure. After generation of a new structure, the structure is updated to avoid generated structures without introducing a bias over the remaining set of structures. This ensures non-redundancy of the sampling algorithm, in turn allowing faster access to higher free-energy locally optimal intermediate states. The data-structure is also easily portable and can be incorporated into most of current recursive sampling algorithms.
Finally, the dynamic programming scheme underlying RNANR can be richly parametrized, allowing to limit the searching space without inducing supplementary complexity costs.

Results. Our tests on both real and random RNAs reveal that RNANR generates more unique structures in a given time than its competitors, and allows a deeper exploration of RNA kinetics landscapes. Thanks to its non-redundancy, our software generates unique samples more that 2 times faster than RNAlocmin or RNAlocopt. This difference becomes even more apparent for bigger sizes of unique samples due to smaller probability of finding new unique samples by redundant sampling. We also performed a comparison between RNANR and main competitors - RNASLOpt, RNAlocopt, RNAlocmin and RNAsubopt - over a dataset of artificial riboswitches, RNAs designed to exhibit a bistable behavior. Each of 250 sequences is 100 nt long and presents two low free-energy states that are distant by at least 20 base-pairs. In 70% of the instances, structure sets generated using RNANR provide more accurate picture of RNA kinetics, as measured using standard numerical integration. This demonstrates that the deeper exploration of RNA kinetic landscapes allows to retrieve important transient secondary structures, leading to approximated kinetic landscapes that are closer to the reality.

On a theoretical level, the non-redundant sampling can be employed to estimate various statistics
in the Boltzmann ensemble. While naive estimator based on the expectation is in this case biased,
we developed a new non-biased estimator that converges faster than naive estimator used on non-redundant sample. We compared the performance of both estimators on samples of 10000 secondary structures for each of 365 sequences from RFAM families RF00001, RF00005, RF00061, RF00174, RF01071 and RF01731 [9], while the statistics of interest was presence of a specific base-pair. The non-redundant estimator performed better in $\sim$ 80%, mostly for shorter sequences
and samples covering bigger part of Boltzmann ensemble, while %GC did not seem to have an impact of the performance of the estimators.

Perspectives.
In future extensions, we plan to include simple pseudoknots and kissing hairpins
into our sampling algorithms. While many secondary structures are pseudoknot-less, a number of
them is stabilized by pseudoknots, as shown by 250 pseudoknotted structures in Pseudobase++ [10],
making them important for both RNA thermodynamics and kinetics. While complicated pseudoknots need high-complexity algorithms to be detected induce high energy penalties making them unlikely, simpler pseudoknot classes do not suffer from similar problems. For this reasons, it is desirable for our algorithm to be able to detect them and we are investigating parsimonious strategies
to determine an optimal tradeoff between expressivity and computational demands.

References

Juraj Michálik, Hélène Touzet, and Yann Ponty. Effcient approximations of rna kinetics landscape using non-redundant sampling. Bioinformatics (Oxford, England), 33:i283–i292, July 2017.

Azadeh Saffarian, Mathieu Giraud, Antoine De Monte, and Hélène Touzet. RNA locally optimal secondary structures. Journal of Computational Biology, 19(10):1120–1133, 2012.

Yuan Li and Shaojie Zhang. Finding stable local optimal RNA secondary structures. Bioinformatics, 27(21):2994–3001, 2011.

Ronny Lorenz, Stephan H Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA package 2.0. Algorithms for Molecular Biology, 6(1):26, 2011.

William A Lorenz and Peter Clote. Computing the partition function for kinetically trapped RNA secondary structures. PLoS One, 6(1):e16178, 2011.

Marcel Kucharík, Ivo L Hofacker, Peter F Stadler, and Jing Qin. Basin Hopping Graph: a computational framework to characterize RNA folding landscapes. Bioinformatics, 30(14):2009–2017, 2014.

Jérôme Waldisp´uhl and Peter Clote. Computing the partition function and sampling for saturated secondary structures of RNA, with respect to the turner energy model. Journal of Computational Biology, 14(2):190–215, 2007.

David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proceedings of the National Academy of Sciences of the United States of America, 101(19):7287–7292, 2004.

Ioanna Kalvari, Joanna Argasinska, Natalia Quinones-Olvera, Eric P Nawrocki, Elena Rivas, Sean R Eddy, Alex Bateman, Robert D Finn, and Anton I Petrov. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic acids research, 46(D1):D335–D342, 2017.

Michela Taufer, Abel Licon, Roberto Araiza, David Mireles, FHD Van Batenburg, Alexander P Gultyaev, and Ming-Ying Leung. Pseudobase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. Nucleic acids research, 37(suppl 1):D127–D135, 2008.

Keywords: RNA, kinetics, thermodynamic model, combinatorial algorithm, dynamic programming, statistical learning

A voltage-gating mechanism in plant aquaporins

Robin Mom *† [1], Jean-Stéphane Venisse [1], Caroline Savel [1], Boris Fumanal [1], Aurélie Gousset [1], Marc Vandame [1], Patricia Drevet [1], Gilles Pétel [1], Agilio Padua [2], Philippe Label [1], Daniel Auguin‡ [3]

[1] Laboratoire de Physique et Physiologie Intégratives de l'Arbre en environnement Fluctuant - Clermont Auvergne (PIAF) – Université Clermont Auvergne : UMRA547, Institut national de la recherche agronomique [Auvergne/Rhône-Alpes] : UMRA547 – INRA Site de Crouël 234 / avenue du Brézet / 63100 Clermont-Ferrand - France, France

[2] Institut de Chimie de Clermont-Ferrand - Clermont Auvergne (ICCF) – SIGMA Clermont, Université Clermont Auvergne : UMR6296, Centre National de la Recherche Scientifique : UMR6296 – 24 Avenue des Landais / 63177 Aubière Cedex, France

[3] Laboratoire de Biologie des Ligneux et des Grandes Cultures (LBLGC) – Université d'Orléans – rue de Chartres BP6759 45067 Orléans CEDEX 2, France

Water is essential for any living organism as it plays a crucial role in metabolism and regulation of cells homeostasis. In plants especially, growth and development rely upon tight regulation of its movements. Water diffusion across biological membranes is facilitated in particular by aquaporins (AQP). They offer a very flexible and rapid way to the plant to regulate transcellular water flows and cope with an ever-changing hydric environment. AQP are found across all kingdoms of life but present the highest diversity in fish, mammals and higher plants in which duplication coupled with horizontal gene transfer events has led to various neo-functionalized isoforms with different solute selectivity, different gating mechanisms or different expression in time and space. Notwithstanding the pore residues are very conserved in the AQP family because of their crucial role in the transport activity of the protein. Among them, four constitute the aromatic/arginine (ar/R) constriction which corresponds to the narrowest part of the channel and of the pore's region where protein-water interactions are dominant. Hence, the ar/R constriction is believed to play a prominent role in the selectivity of the pore. The arginine is particularly important for protein-solute interactions and AQP share at least 90 % identity for this amino acid. Recently, molecular dynamics studies led on human AQP pointed out a putative new function of this constriction site involving in a new gating mechanism controlled by voltage. Indeed, because of its positively charged guanidinium group, this arginine would act as an electrostatic field sensor, oscillating between an open up and a closed down conformational state. Here, through molecular dynamics simulations of plant AQP SoPIP2 (PDB 1z98), we observed for the first time this voltage-gating in the plant realm. Because of the extremely conserved arginine involved, we hypothesize that this mechanism could constitute a very basic feature of AQP, working as a safety gating sensitive to sudden stresses and allowing a very quick response of the cell on the scale of the nanosecond. However, to elucidate the relevance of voltage sensing

[*] Speaker

[†] Corresponding author: robin.mom@etu.uca.fr

[‡] Corresponding author: auguin@univ-orleans.fr

in AQP, more experiments need to be carried out, especially in vivo, as water transport through AQP has yet not been linked to voltage other than through molecular dynamics simulations.

Keywords: aquaporins, molecular dynamics, voltage, gating, SoPIP2

Proteins and their multiple interactions

Chloé Dequeker *† 1, Elodie Laine‡ 2, Alessandra Carbone§ 3,4

[1] Laboratoire de Biologie Computationnelle et Quantitative (LCQB) – Centre National de la Recherche Scientifique : UMR7238 – Biologie Computationnelle et Quantitative UMR 7238 CNRS-Université Pierre et Marie Curie 4 place Jussieu, 75005 Paris, France, France

[2] Laboratoire de Biologie Computationnelle et Quantitative (LCQB) – Sorbonne Universités, Université Pierre et Marie Curie (UPMC) - Paris VI, CNRS : UMR7238, IBPS – 4 Place Jussieu, 75005 Paris, France

[3] Institut Universitaire de France (IUF) – Ministère de l'Enseignement Supérieur et de la Recherche Scientifique – France

[4] Laboratoire de Biologie Computationnelle et Quantitative (LCQB) – Sorbonne Université UPMC Paris VI, Centre National de la Recherche Scientifique - CNRS – France

Proteins are main actors in biological processes and a detailed description of their interactions with other proteins, nucleic acids and small molecules, is expected to provide direct information on these processes and on the way to interfere with them. A comprehensive view of a protein's cellular partners is given by its interaction network. Our knowledge of interaction networks is largely incomplete, as the experimental assessment of all possible interactions of a protein is very challenging [1,2]. But, if the finer residue resolution of the interactions is provided, the network may also describe the way in which the protein interacts with other molecules.

In living cells, proteins are expected to engage in multiple interactions taking place either at the same moment, on different sites of the protein surface, or at different moments, possibly on a shared site. The precise prediction of these multiple sites would provide useful information on the number of interactors in the protein network. Moreover, learning about the specific evolutionary, physico-chemical and geometrical properties of the interaction sites would be useful for inferring protein functional activities and for predicting potential interactions established by the protein during its lifetime or upon mutations.

In the present study, we combine information coming from evolutionary sequence conservation, statistically derived physico-chemical properties expected at the interface, and local geometry of the protein surface with physics based properties coming from docking analysis to demonstrate how these features can help to identify complementary prediction strategies and to increase the accuracy of existing approaches. For this purpose, we make use of a Complete Cross-Docking (CCD) performed on 2246 human proteins involved in the Muscular Dystrophy (HCMD2 project) from which we extracted 262 proteins on which docking behaviour was described.

Binding site predictions are obtained with the dynJET2algorithm, an updated version of the existing tool JET2 [3, 4], integrating the four features in four different scores. These different

*Speaker

†Corresponding author: chloe.dequeker@upmc.fr

‡Corresponding author: elodie.laine@upmc.fr

§Corresponding author: alessandra.carbone@lip6.fr scores yield several distinct or partially overlapping interaction patches, thus accounting for the multiplicity of interactions a protein may establish during its lifetime. NIP is the normalised form of the Interface Propensity score IP , that reflects the propensity of a residue to be found at the interface. In order to compare IP scores among proteins, we normalise it, as done in [5]: a positive NIP value indicates that the residue I is favour to occur at potential binding sites, and a negative NIP value indicates that it is disfavoured. The predictions thus obtained allow us to single out interacting partners, when combining them to docking information.

We analysed a dataset of 262 proteins for which at least one structure bound to a partner was deposited in the PDB. They were extracted from the larger set of 2 246 proteins used to perform a complete cross-docking study [5], leading to 4 years of computation on the World Community Grid (WCG, www.worldcommunitygrid.org [1]). Starting from the observation [6] that functional interfaces are conserved across closely related homologs, we retrieved all interacting surfaces described by complexes in the PDB involving either a protein from the dataset or a close homolog ($\geq 90\%$ sequence identity). By coupling these interacting surfaces, we were able to define experimental interaction regions (used by several partners) and interaction sites (used by a single partner) for each protein, recovering as much information as possible on the multiple interactions that the protein might have in the cell. We could then compare our predictions to the enriched set of potential interaction regions for a protein surface and showed that dynJET2 manages to accurately predict both interaction sites and interaction regions. As a consequence of our effort in retrieving as much information as possible on protein interactions from homologous structures in the PDB, we demonstrate that the evaluation of protein-protein interface algorithms cannot be correctly assessed by relying on one single complex for a given protein.

References:

: E. L. Huttlin, L. Ting, R. J. Bruckner, F. Gebreab, M. P. Gygi, J. Szpyt, S. Tam, G. Zarraga, G. Colby, K. Baltier, R. Dong, V. Guarani, L. P. Vaites, A. Ordureau, R. Rad, B. K. Erickson, M. Wuhr,
J. Chick, B. Zhai, D. Kolippakkam, J. Mintseris, R. A. Obar, T. Harris, S. Artavanis-Tsakonas, M. E. Sowa, P. De Camilli, J. A. Paulo, J. W. Harper,and S. P. Gygi. The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell,162(2):425–440, Jul 2015.

: T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A. R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Trigg, J. C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A. L. Barabasi, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, and M. Vidal. A proteome-scale map of the human interactome network. Cell, 159(5):1212–1226, Nov 2014.

: Elodie Laine and Alessandra Carbone. Local geometry and evolutionary conservation of protein surfaces reveal the multiple recognition patches in protein-protein interactions. PLOS Computational Biology, 11(12):1–32, 12 2015.

: H. Ripoche, E. Laine, N. Ceres, and A. Carbone. JET2 Viewer: a database of predicted multiple, possibly overlapping, protein-protein interaction sites for PDB structures. Nucleic Acids Res., 45(7):4278, Apr 2017.

: A. Lopes, S. Sacquin-Mora, V. Dimitrova, E. Laine, Y. Ponty, and A. Carbone. Protein-Protein Interactions in a Crowded Environment: an Analysis via Cross-Docking Simulations and Evolutionary Information. PLoS Comput. Biol., 9(12):e1003369, 2013.

: Buyong Ma, Tal Elkayam, Haim Wolfson, and Ruth Nussinov. Protein–protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. PNAS, 100(10):5772–5777, May 2003.

Keywords: protein, interactions, protein, protein

Promiscuous binding site: a frequent material of proteins

Natacha Cerisier *† 1, Manon Réau , Quentin Bayard , Leslie Regad 1, Michel Petitjean 1, Anne Badel 1, Anne-Claude Camproux‡ 1

1 Molécules thérapeutiques In silico (MTi), INSERM UMR-S973, Université Paris Diderot (MTi) – Université Paris Diderot - Paris 7 – 35 rue Hélène Brion, 75013 Paris Cedex, France, France

Over the last 50 years, drug development processes have been based on the "one drug - one target" paradigm. The primary goal of drug discovery has been to design and deliver selective compounds against individual biological targets. Nowadays, it is well-known that a drug may be involved in different disease functions and interact with more than one target which defines the drug promiscuity [Haupt at al. 2013]. Consequently, the drug promiscuity is receiving a lot of attention. According to a recent study [Jalencas et al. 2013], only 15% of the drugs from several databases interact with one single target whereas over 50% of them interact with more than five targets. The consequences of drug promiscuity can be either beneficial or undesirable. Beneficial outcomes include its possible application to new diseases and the time and money saved on preclinical tests. Amongst the undesirable outcomes, the promiscuous drugs can interact with off-targets, resulting in adverse drug reactions, harmful side effects and adverse polypharmacology. When drugs are used to target several disease-related pathways, polypharmacology [Lavecchia et al. 2016] is now recognized as an increasingly important aspect of drug design. Drug polypharmacology is analyzed by the physicochemical properties and fragment composition of the drug, but also in terms of the protein family and distant binding site similarities of the main drug's target. Global structure and binding site similarity were shown to have greater influence on the drug promiscuity than the aforementioned drug properties, such as hydrophobicity, molecular weight or ligand flexibility [Haupt et al. 2013]. Jalencas et al, (2013) speculated that the levels of polypharmacology observed in current drugs may be a quiescent signature of evolution itself. Early biological organisms could have developed adaptation mechanisms such as increasing protein plasticity to favor chemical promiscuity, eventually leading to functional promiscuity. Kufareva et al. (2012) confirmed that the plasticity of the binding site is important in protein–ligand interactions by constructing a Pocketome encyclopedia, a collection of druggable binding sites. Gao et al. (2013) gathered 20 000 ligand-bound pockets resulting in only 1 000 representative pockets of which 1/3 were considered as promiscuous, meaning able to interact with multiple, chemically different ligands. Bajorath et al. (2016) explore the promiscuity from the target perspective with a ChEMBL20 dataset and conclude that the majority of targets interact with structurally diverse compounds.

These proteochemometric approaches take advantage of the increasing number of 3D structures in complexed form to improve knowledge of the protein-ligand interactions. They emphasize the interest to explore more the target binding site space. To our knowledge, there is no exhaustive

[*] Speaker

[†] Corresponding author: natacha.cerisier@univ-paris-diderot.fr

[‡] Corresponding author: anne.claude.camproux@univ-paris-diderot.fr

work dedicated to quantifying and studying promiscuous binding sites within protein families. Here we investigate the binding site promiscuity frequency of targets and explore why some binding sites are able to bind to one specific drug. We quantify the promiscuity of druggable binding sites using the Mother Of All Databases (MOAD, [Ahmed et al 2015]). MOAD is one of the largest databases that provides more than 23 000 protein-ligand complexes, with only high-quality resolution structures extracted from the PDB (X-ray structures, less than 2.5 Å). From the MOAD, we focus on 3831drug-like valid ligands and 8669 corresponding protein-ligand interactions. We concentrated on druggable binding sites, able to accommodate drug-like molecules (orally bioavailable small drugs that have an optimal profile of physicochemical properties, in terms of Absorption, Distribution, Metabolism, Excretion and Toxicity (ADME-Tox) as defined by Lipinski in 1997). Indeed, the druggability is a very important point in drug design [Hussein et al, 2017]. We choose to describe a binding site by all the overlapped pockets (cavities) estimated by proximity to a valid drug-like ligand from MOAD complexes. A cavity is defined as a set of atoms that are in direct contact with a ligand in a complex structure. This direct contact is quantified by a given threshold to the ligand, of 5.5 Å, as in Borrel et al (2015). Thus, a binding site from one protein can be described by the corresponding superset of pockets. We studied the frequency and representativeness of promiscuous binding sites in different MOAD protein families. This study confirms that promiscuous binding sites are observed with high frequency, about 60%, in diverse protein families, which underlines their possible impact on multiple drug-target interaction modeling. It confirms that binding site promiscuity analysis and detection can contribute to drug repositioning or off-target detection. Then, we studied the diversity and characteristics of pockets and ligands corresponding to different binding site promiscuity, in terms of their geometrical and physico-chemical properties and provided prediction of these promiscuous and highly promiscuous binding sites.

Keywords: Promiscuous, proteins, polypharmacology, bioinformatic, MOAD, pocket

Mass spectrometry data-independent acquisition (DIA/SWATH-MS) management and processing with myProMS web server

Marine Le Picard [1], Alexandre Sta [2], Bérangère Lombard [1], Vanessa Masson [1], Guillaume Arras [1], Stéphane Liva [2], Florent Dingli [1], Damarys Loew [1], Emmanuel Barillot [2], Patrick Poullet *† [2]

[1] Proteomics and Mass Spectrometry Laboratory (LSMP) – Institut Curie, PSL Research University – 26 rue d'Ulm 75240 Paris cedex 05, France

[2] Bioinformatics Platform - INSERM U900 - Bioinformatics and Computational Systems Biology of Cancer – Institut Curie, Inserm U900, Mines Paris Tech, PSL Research University – 26 rue d'Ulm 75240 Paris cedex 05, France

Cell-wide profiling of differentially expressed proteins under many perturbation conditions, or between normal and disease states (e.g. cancer) is becoming a mandatory counterpart of genomic and transcriptomic approaches to capture the complexity of the biological processes at work. Mass spectrometry (MS) has become the prominent technology allowing the identification and quantification of thousands of proteins and post-translational modifications (PTMs) from biological samples of interest. However, traditional MS data-acquisition mode used in Discovery proteomics, Data-Dependent Acquisition (DDA), suffers from sampling issues impairing quantification precision and reproducibility. During the past few years, a new acquisition method, Data-Independent Acquisition (DIA or SWATH-MS), has emerged that provides proteome-scale quantification coverage with much higher accuracy than DDA (Gillet, 2012). This method requires faster mass spectrometers but is becoming within reach of an increasing number of MS facilities as they acquire compatible instruments. The major drawback of DIA is that the data generated are complex and require multiple steps of data processing and statistical analysis to extract meaningful biological information. Multiple commercial and open-source solutions have been developed or adapted to process DIA data such as the Trans-Proteomic Pipeline (TPP) (Keller, 2005), skyline (MacLean, 2010), the Galaxy-P project (http://usegalaxyp.org), OpenMS (Sturm, 2008) or more recently OpenSWATH (R΄ost, 2014), DIA-Umpire (Tsou, 2015), PeakView (ABSciex) or SpectronautTM Pulsar (Biognosys). Unfortunately, the free tools available are difficult to set up for most MS facilities which have limited (bio)informatics resources. In addition, DIA-processing tools are much too technology-oriented and lack data integration functionalities. These limitations hamper not only their easy deployment for evaluation but also their use in the context of large-scale studies.

In order to facilitate MS data management and analysis, we have initiated the myProMS project in 2004 as a collaboration between the Institut Curie Bioinformatics and MS platforms (Poullet, 2007). The aim of this project is to develop a comprehensive bioinformatics solution allowing MS-based proteomics data centralization, integration, analysis and easy access by both

*Speaker

†Corresponding author: patrick.poullet@curie.fr

MS facility members and research collaborators through a shared web server. Over the years, myProMS has evolved into a robust environment for conventional DDA data analysis including support for major MS search engines and sample labelling strategies, peptide/protein identification validation, peptide/protein/PTM quantification, exploratory (PCA, clustering,...) and functional (Gene Ontology, pathway,...) analyses. Recently, we have upgraded the server to support DIA/SWATH-MS data integration and processing. myProMS relies on the TPP for spectral library generation and on the Open(MS/SWATH) workflows for peptide identification and transition quantification; both tools being fully integrated to the server. In addition, results from PeakView and SpectronautTM can be seamlessly imported for further processing including protein differential analysis performed using the R package MSstats (Choi, 2014).

Thus, myProMS constitutes a unique environment for discovery proteomics where high-throughput proteomic data generated from both classical DDA and DIA/SWATH-MS data are unified and can be seamlessly processed, mined or exported for further analysis such as integration with other -omic data.

myProMS is the outcome of a successful 15-year collaboration between a bioinformatics platform and a MS facility. Such a long-term partnership is rare enough to be mentioned. It is the basis of the successful continuous adaptation of the tool to match the fast evolution of proteomic MS technology.

myProMS is freely available on demand as a Docker image. Contact: myproms@curie.fr.

References

Choi M. et al. (2014) MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics, 30(17):2524-6.

Gillet L.C. et al. (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol. Cell. Proteomics 11(6), O111.016717.

Keller A. et al. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol. Syst. Biol., 1:2005.0017.

MacLean B. et al. (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics, 26(7):966-8.

Poullet P. et al. (2007) myProMS, a web server for management and validation of mass spectrometry-based proteomic data. Proteomics, 7(15):2553-615.

R´ost L.H. et al. (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nature Biotechnology, 32(3):219-23.

Sturm M. et al. (2008) OpenMS – an Open-Source Software Framework for Mass Spectrometry, BMC Bioinformatics, 9: 163.

Tsou C.C. et al. (2015) DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Methods, 12(3):258-64.

Keywords: Mass spectrometry, Proteomics, Data Independent Acquisition, DIA/SWATH MS, MSstats, Data management, Web server

Advanced molecular dynamics simulations for understanding the functions and dysfunctions of the CFTR channel

Ahmad Elbahnsi * [1], Fabio Pietrucci [1], Brice Hoffmann [1], Pierre Lehn [2], Jean-Luc Decout [3], Jean-Paul Mornon [1], Isabelle Callebaut

[1] Institut de minéralogie, de physique des matériaux et de cosmochimie (IMPMC) – Institut de recherche pour le développement [IRD] : UR206, Université Pierre et Marie Curie (UPMC) - Paris VI, CNRS : UMR7590, Muséum National d'Histoire Naturelle (MNHN) – Tour 23 - Barre 22-23 - 4e étage - BC 115 4 place Jussieu 75252 PARIS, France

[2] Unité INSERM 1078, SFR ScInBioS – Université de Bretagne Occidentale (UBO) – 46 rue Félix Le Dantec, CS51819, 29218 Brest, France

[3] Département de pharmacochimie moléculaire (DPM) – CNRS : UMR5063, Université Joseph Fourier - Grenoble I – Batiment E André Rassat 470 rue de la chimie - BP 53 38041 GRENOBLE Cedex 9, France

Mutations of the Cystic Fibrosis Transmembrane conductance Regulator (CFTR) anion channel lead to the inherited disease Cystic Fibrosis. Thus, CFTR represents an important target for drug discovery, with challenging issues related to the development of modulators specific for each class of mutations.

Significant progress has been achieved last year with the first 3D structures of the full-length CFTR, solved using cryo-electron microscopy. These, in both dephosphorylated, ATP-free and phosphorylated, ATP-bound conformations, represent inactive states of the channel. Other, active conformations of the CFTR channel have been proposed through modeling in previous studies, based on ABC exporter templates. Among these are our models of human CFTR in open and closed conformations, which were supported by various experimental studies and are also in good agreement with the now published cryo-EM experimental structures. Despite these crucial structural data, we still need to further understand the transition mechanisms between the different conformational states, evaluate the impact of mutations on these processes and appreciate how the 3D structure information can be used to rationalize drug binding and design. In order to address these issues, we have set up metadynamics simulations as a tool to unveil new insights into the channel dynamics and thermodynamics. This approach aims at i) enhancing rare events and thus overcoming standard molecular dynamics limitations and ii) obtaining the free energy landscape connecting two distinct states of the channel. Metadynamics simulations were first performed to analyze the transition path linking our open and closed models, showing in particular that the closed form is energetically favored over the open form. The metadynamics setup is now further implemented to characterize CFTR conformational transitions that are taking place between inactive (experimental) and active (our models) states. We exploited the different CFTR conformers to characterize the impact of CFTR mutations and understand the molecular mechanisms of modulation by identifying modulator binding sites.

*Speaker

Keywords: metadynamics, CFTR, membrane protein

Data science

mmquant and mmannot: How to handle multiple-mapping reads in (s)RNA-Seq

Matthias Zytnicki *† 1, Christine Gaspin 1

1 MIAT – Institut national de la recherche agronomique (INRA) – France

mmquant and mmannot: How to handle multiple-mapping reads in (s)RNA-Seq

Matthias Zytnicki, Christine Gaspin

INRA MIAT, Toulouse

Introduction

RNA-Seq and small RNA-Seq (sRNA-Seq) are currently used routinely, and they provide accurate information on gene and small-RNA transcription. However, the methods cannot accurately estimate duplicated transcript expression. Several strategies have been previously used (drop duplicated transcripts, distribute uniformly the reads, or estimate expression), but all of them provide biased results. We provide here two tools, called mmquant (published in [6]) and mmannot (submitted for publication), for computing expression of genes and small RNAs respectively, including duplicated elements.
mmquant is available at https://bitbucket.org/mzytnicki/multi-mapping-counter, and mmannot is available at https://sourcesup.renater.fr/wiki/mmannot.

mmquant: a strategy for the gene quantification including multi-mapping reads

In general, RNA-Seq quantification reads a set of reads files (one file per sample) and a annotation file that lists the set of known genes. It produces a count table, where a row is a gene, a column is a sample, and each cell provides the number of reads matching a gene in a sample. The aim is usually to find differentially expressed genes, i.e. the set of genes that are more, or less, expressed in a subset of the samples when compared to the other samples.

So far, three strategies are used when a read may map at several positions:
- a ”unique” method: discard multi-mapping reads,
- a ”random” method: use a random hit,
- a ”ratio” method: weight each hit (if a read maps n times, each hit counts for 1/n).

mmquant implements an other strategy, firstly presented in [4]. If a read maps at different positions, mmquant detects that the corresponding genes are duplicated; it merges the genes

[*] Speaker
[†] Corresponding author: Matthias.Zytnicki@inra.fr

and creates a ”merged gene” feature, which appears as a new line in the count table. As a result, the differentially expression test can be performed similarly on the regular genes and on the merged genes. mmquant is a drop-in replacement of the widely used tools htseq-count [2] and featureCounts [3] that handle multi-mapping reads in an unbiased way.

The tool supports paired-end reads, and checks that both ends may match the same transcript, in a way that is consistent with the sequencing strategy (forward-reverse, reverse-forward, etc.). The fragments (i.e. the pairs of reads) are then counted for quantification.

We tested our method on several data sets on different species, but for space reasons we focus on the human data set, taken from [1]. Briefly, this study uses RNA-Seq of human brain to find genes that are differentially expressed
in individuals diagnosed with bipolar disorder. Admittedly, this dataset is challenging because duplicated genes are known to play a major role in human brain.

Strikingly, the p-values obtained with the three different quantification strategies show a great variability. htseq-count, featureCounts and mmquant (excluding merged genes) gave 734, 835 and 763 differentially expressed genes respectively. Most of the differences comes from the way reads are assigned to the genes.

mmquant found that 5-6% of the reads where multi-mapped and could be attributed to several genes. As a consequence, it found 254 additional differentially expressed merged genes, involving 516 new genes. Note that one fourth of the differentially expressed genes is merged.

We then considered the 33 merged genes with adjusted p-value $<$ 1%, which represented very good candidates. These merged genes included 75 genes that were not detected otherwise (neither by htseq-count nor featureCounts, nor in the non-merged genes found by mmquant). This gene list includes new excellent candidates with putative links to bipolar disorder, including ADK, GTF2I, hnRNP-A1, HTRA2, PKD1 and RERE, which have been linked to various brain-related diseases. Some of these genes have complex regulation systems in cis: ADK and HTRA2 contain overlapping processed pseudogenes and antisense transcripts or genes, and mmquant merges these annotations on the fly. Other genes, like GTF2I, hnRNP-A1, PKD1, and RERE, are duplicated genes, or have produced a pseudo-gene in another locus. It is out of the scope of this study to validate these genes, but we would like to emphasize that, because these genes are duplicated, or overlap with other genes, they have been removed from the standard analysis.

Concerning time, featureCounts is the fastest tool, taking 8-11min per sample; mmquant is second with 21-29 min (+1-3 min if the reads are not sorted); htseq-count, written in Python, takes 4h15min-5h29min. mmquant is slower than featureCounts because it has to store (and look up) all the reads that have been mapped several times. We obtained this results allocating one thread per BAM file, but featureCounts can be further accelerated by allocating more than one thread per input file, whereas mmquant and htseq-count cannot.

mmannot: a strategy to quantify repetitive small non coding RNAs

Small non coding RNAs gather a very wide collection of classes, such as microRNAs, tRNA-derived fragments, small nucleolar RNAs and small nuclear RNAs, to name a few. As usual in RNA-seq studies, the sequencing step is followed by a feature quantification step: when a genome is available, the reads are aligned to the genome, and the corresponding features are quantified.

The sRNA classes are then quantified by counting the number of reads co-localizing with the

members of each class. Although simple and widely used, this strategy does not work in several ambiguous cases.
- A read maps at several loci: if two different regions of the genome are identical (usually after a genome duplication), a read may map equally well at different locations.
- Frequently in sRNA-Seq, two different annotations overlap in the genome and a hit (i.e. a read mapping) overlaps both: In this case the hit may be attributed to either annotation
- A hit co-localize two different annotations, even though the annotations do not overlap themselves: The hit is usually at the frontier of the annotations.
The first source of ambiguity arise often, because some sRNAs are known to co-localize with duplicated regions (such as piRNAs or siRNAs), or to be included into duplicated genes (miRNAs and tRFs).

mmannot implements a strategy similar to mmquant, that compares all the reads that map at several positions, and their annotations when available. In many cases, all the hits co-localize with the same feature annotation (a duplicated miRNA or a duplicated gene, for instance). When different annotations exist for a given read, we propose to merge existing features and provide the counts for the merged features.

A configuration file is required to select the annotations that should be quantified. The configuration file ranks the annotation by order of priority, but ties are accepted. Using the exon annotation usually provided in the annotation file, mmannot automatically extracts introns, coding sequences, 5' and 3' untranslated regions (UTRs), down- and upstream regions of the features selected by the users (e.g. coding genes, non-coding genes), and adds them in the in-memory annotation dataset.
The user can also specify a strand orientation of the read with respect to the annotation (collinear or antisense).

The process of read annotation proceeds in two steps. The first step aims at finding the matching annotations of a given hit. If a hit matches several different annotations (e.g. miRNA and intron), the annotation with highest priority is kept (here, the miRNA). If several annotations have the same priority, then all of them are kept and the hit is already ambiguous.

The last step resolves the ambiguities. If a read maps uniquely, with no ambiguity, the count of the corresponding annotation is incremented. If a read maps at different locations, but all the hits match the same annotation, we declare that the read is rescued and the corresponding annotation count is also incremented. Likewise, if a read maps only one annotation and intergenic regions, we consider that the read belongs to this annotation, and the read is rescued. If the read or a hit overlaps several annotations, the annotations with highest priority are kept. If there is only one annotation with highest priority, the read or the hit is not ambiguous. Otherwise, there is an ambiguity: we create a new annotation type, called a merged annotation, which is the concatenation of the matching annotations, and its count is incremented. For instance, if a read maps to a 3'UTR and a miRNA, the count of the 3'UTR-miRNA will be incremented. After having scanned all the reads, the quantification table is produced.

We compared the results of the three strategies mentioned previously with the strategy implemented in mmannot. We used datasets of experiments already published, covering several eukaryotic organisms, but we will focus on an *Arabidopsis thaliana* data set [5]. Our aim was to show how each strategy impacts the results of quantification in each class.
We found that, the "unique" method, arguably the most used one, provides very biased results in terms of representative percentage of the class due the strategy used. This strategy annotates around 40% of reads as miRNAs, whereas the other strategies, when considering multi-mapping reads, annotate only around 20% of the reads in the miRNA class. For all datasets, using multi-

mapping strategies increases considerably the percentage of annotated reads, showing that the repertoire of expressed regions is largely associated to repeated regions in all genomes. Some multi-mapping reads may have hits that do not co-localize with any annotation. These reads may be unannotated by the "random" strategy, and the associated weight in the "ratio" strategy is lost. As a consequence, some of these reads are not quantified, resulting in less annotated reads.

We focused on ambiguous reads to analyze the origin of reads annotated as ambiguous in that organism. Most of them involve downstream regions, upstream regions, or introns, and they are probably intergenic duplicated regions. Then, the most frequent ambiguous annotations involve a transcribed region within the downstream, upstream, or intron regions. These elements might be produced by some unannotated regions that target genes, e.g. siRNAs, or belong to a non-functional genomic duplication. Interestingly, the most frequent ambiguous annotation involving two transcribed regions is miRNA-gene (-). In plants, a miRNA and its target may be 100% identical, and thus pose a real problem to the annotation.
We studied the pairs of miRNAs–genes that were involved in this class, with at least 100 reads supporting this association.
We found well-known miRNAs and their targets at mapped loci: miR156/miR157 with SPL, miR163 with PXMT1, miR171 with ATHAM, miR400 with PPR1, miR403 with Ago2 and miR824 with AGL. Note that these reads cannot be correctly annotated by any other method, and the expression of the miRNAs are thus under-estimated.

Conclusion

Transcript quantification is an essential step of many RNA-Seq analyses. Yet, the assumption used by the quantification tools is not always fully understood, especially concerning multi-mapping reads. With mmquant and mmannot, we provide simple tools, that include these reads in the quantification step, with no assumption on the read distribution. We hope that that these tools could be used as a drop-in replacement of previous tools, and that part of the genomic "dark matter" will be at last explored.

References

N Akula, J Barb, X Jiang, J Wendland, KH Choi, SK Sen, L Hou, DTW Chen, G Laje, K Johnson, BK Lipska, JE Kleinman,
H Corrada-Bravo, S Detera-Wadleigh, PJ Munson, and FJ McMahon. *RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder.* Molecular psychiatry, 19:1179–1185, 2014.

Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. *Htseq–a Python framework to work with high-throughput sequencing data.* Bioinformatics, 15:166–169, 2015.

Yang Liao, Gordon Keith Smyth, and Wei Shi. *featurecounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 30:923–930, 2014.

Christelle Robert and Mick Watson. *Errors in RNA-Seq quantification affect genes of relevance to human disease.* Genome Biology,16:177, 2015.

Elena A Vidal, Tomás C Moyano, Gabriel Krouk, Manpreet S Katari, Milos Tanurdzic, W Richard McCombie, Gloria M Coruzzi, and Rodrigo A Gutiérrez. *Integrated RNA-seq and sRNA-seq analysis identifies novel nitrate-responsive genes in* Arabidopsis thaliana *roots.* BMC Genomics, 14:701, 2013.

Matthias Zytnicki. *mmquant: how to count multi-mapping reads?* BMC Bioinformatics, 18:411, 2017.

Keywords: RNA, Seq, sRNA, Seq, NGS

RNA complex prediction as a constrained maximum clique problem

Audrey Legendre * [1], Eric Angel [1], Fariza Tahi† [1]

[1] IBISC, Univ Evry, Université Paris Saclay (IBISC) – Université d'Evry-Val d'Essonne : EA4526 – 23, Bd de France 91034 - EVRY, France

Introduction

RNAs can interact and form complexes, that have specific biological roles. The prediction of their secondary structure is a first step towards the identification of their 3D structure. There exists many tools for complexes composed of two RNAs but there are very few for those composed of more than two RNAs. This is a difficult task, especially when considering pseudoknots, crossing interactions and zig-zags.

The first proposed tool for RNA complex prediction was MultiRNAFold [4] which connects in some order the RNA strands to each others and then computes their energy. The NUPACK package [13], proposed later, includes a software to predict the minimum free energy of an RNA complex by computing the partition function. In [12], the prediction of RNA complexes is formalized using the combinatorial optimization problem called Pegs and Rubber Bands [2] and an approximation algorithm is proposed. Then the tools NanoFolder [6] and HyperFold [7] for RNA complex prediction were developed.

All the tools and algorithms presented above used thermodynamic models based on minimum free energy computation. However, the real structure of an RNA is not always the structure of minimum free energy but a structure close to this one. And this applies to the prediction of RNA complex structures. Hence, to be able to generate sub-optimal solutions in RNA complex prediction is an important issue. To our knowledge, only NUPACK provides sub-optimal solutions.

We propose here an original approach and a tool for RNA complex prediction including pseudoknots, crossing interactions and zig-zags.

We formulate the problem of predicting RNA complexes as a combinatorial optimization problem where we must find the best combination (according to a free energy minimization criteria) of predicted RNA secondary structures and RNA-RNA interactions. The secondary structures and interactions are organized into a graph, such that the problem becomes a constrained Maximum Clique Problem (MCP).

We propose an heuristic based on local search [1] and tabu search [8] for this problem and developed a tool called RCPred based on this heuristic for RNA complex prediction, returning several sub-optimal solutions. The software RCPred and the benchmark dataset are available on the EvryRNA platform (https://evryrna.ibisc.univ-evry.fr/).

Predicting RNA complexes: a constrained MCP

The RNA complex prediction problem can be formalized using a weighted graph *G(V, E)* such that:

[*] Speaker

[†] Corresponding author: fariza.tahi@univ-evry.fr

- V is the vertices set. V is composed of two subsets, the set of vertices representing the secondary structures and the set of vertices representing the interactions. Each vertex $v \in V$ has a weight equals to the free energy associated to the structure or the interaction.
- E is the edges set. An edge exists if and only if two vertices are compatible. Two vertices are not compatible if identical nucleotides are involved in different pairings or if the two vertices are secondary structures involving the same RNA. It allows us to predict any motif of RNA complexes: pseudoknots, crossing interactions or zig-zags.

An RNA complex is a set of secondary structures and interactions which are all mutually compatible. An RNA complex then corresponds to a complete subgraph, called a "clique", where each vertex is linked to all the other vertices. This clique is constrained because for each RNA, there must be exactly one secondary structure vertex involving it, in the clique. In some known complexes, the RNAs are not structured, hence we consider for each RNA a vertex corresponding to an empty secondary structure.

However, a clique only composed of empty secondary structures and no interactions, that we call a weak clique, has no biological meaning. We have therefore a constrained maximum vertex weight clique problem, denoted in the sequel by constrained MCP, where the clique should be composed of exactly one secondary structure per RNA, is not weak and has a minimum free energy.

To each secondary structure and each interaction is associated a free energy corresponding to the weight of the vertices. This energy is returned by the tools used for the prediction of the inputs. The free energy of a complex is approximated by the sum of the weights of the vertices of the clique.

Solving the constrained MCP

Finding the optimum clique for the constrained MCP can be done using a linear program with integer variables but solving it exactly is exponential in time. If sub-optimal solutions are needed, the integer program must be solved several times using additional constraints, increasing the running time. Then an heuristic is needed to provide good solutions efficiently. We propose an adaptation of the heuristic, published in [5], that we call BLS-MCP, based on local search [1] and tabu search [8]. The adaptation of the BLS-MCP method for the constrained MCP is needed to take into account the constraints related to the different kinds of vertices and to avoid the weak cliques.

Before describing our BLS-CMCP algorithm, let us give some useful definitions (illustrated in Figure 1). Let C be a clique of *G(E, V)*. Let *PA* be the set composed of all the interaction vertices excluded from C and connected to all the vertices of C. Let *OM* be the set composed of interaction vertices pairs *(v, u)* (or secondary structures pairs) such that v is excluded from C and is connected to all vertices in C except to vertex u in C. Let *OC* be the set composed of all interaction vertices excluded from C. In the BLS-MCP heuristic, there is no such distinctions between several kinds of vertices. Here, we describe the differences between our algorithm, that we call BLS-CMCP, and BLS-MCP:

- Generation of the initial clique: this process consists in selecting randomly an interaction vertex and then selecting for each RNA a secondary structure vertex that forms a clique. Forming a clique is always possible thanks to the empty secondary structures. In the BLS-MCP, the process consists in selecting randomly a vertex and then to add iteratively vertices if they form a clique, until no more vertex can be added.
- Movement 1: an interaction vertex is selected in *PA* and added into C.
- Movement 2: a vertex pair *(v, u)* is selected in *OM* and v is added to C and u is removed from C.
- Movement 3: an interaction vertex is removed from C.
- Movement 4: an interaction vertex from *OC* is added to the clique C. If they do not form a clique anymore, the secondary structure vertices of the clique are replaced and the interaction

vertices of the clique are removed. In the BLS-MCP, a vertex v from OC is added to C, then the clique is repaired by removing the vertices that do not form a clique.
• Sub-optimal cliques: in BLS-MCP method, no sub-optimal cliques are returned. In BLS-CMCP, any new clique discovered is saved.
• Weak cliques: any movement leading to a weak clique is not considered.

Results
We implemented the BLS-CMCP heuristic for RNA Complex Prediction and obtained the tool RCPred. We compared the results we have obtained with RCPred and with NanoFolder [6], NUPACK [13] and MultiRNAFold [4] on a large set of RNA complexes. For RCPred, we used as inputs for each RNA of a given complex, at most 90 secondary structures with or without pseudoknots recovered from the results of BiokoP [10], pKiss [9] and RNAsubopt [11]. At most 90 interactions between each pair of RNAs were recovered, using RNAsubopt.
We used a dataset of 90 non redundant and experimentally validated RNA complexes of at least 24 nucleotides and at most 968 nucleotides of length, gathered from RNA STRAND database [3].
To evaluate the quality of predicted complexes, we used the sensitivity, the positive predictive value (PPV) and the F1-score, computed as follows:

Sensitivity = TP / (TP+FN)
PPV = TP / (TP+FP)
F1-score = 2 × Sensitivity × PPV / (Sensitivity + PPV)

where TP is the number of true positive base pairs, FN is the number of false negative base pairs, FP is the number of false positive base pairs, and TN is the number of true negative base pairs.

We report the obtained results on Figure 2 and Table 1. We can see that RCPred gives better mean F1-scores, sensititivities and PPV than the other tools and that its F1-score obtained on each complex is in most cases the highest one.

Conclusion

We propose a new method and a tool called RCPred, to predict RNA complexes. Given a set of RNAs, the inputs of RCPred are several possible secondary structures of each RNA and several possible interactions of each pair of RNAs. RCPred returns a set of possible complexes corresponding to the best combinations of the inputs considering the free energy. The inputs can be predicted by one of the many existing tools for RNA secondary structure prediction or RNA-RNA interaction prediction, or given as knowledge by the user.
Our tool can consider non-Watson-Crick base pairs if the energy is known by the user. RCPred is the only tool among the state-of-art that can handle non-Watson-Crick base pairs. RCPred is also the only tool that is able to returning structures including pseudoknots, crossing interactions or zig-zags and sub-optimal solutions. Moreover, some tools of the state-of-art require to order the RNAs before submitting a prediction. The prediction results depending on this order, the user should test all the possible orders but in pratice this is not convenient. RCPred does not depend of any RNA ordering since it is based on a graph representation of the problem.
Each returned complex has a global free energy resulting from the sum of the free energies of the secondary structures and of the interactions composing it. An improvement could therefore to compute in a more sophisticated way the free energy (by adapting for instance the calculation method used in RNAeval [11]) for the returned complexes and reorder them accordingly.
Finally, the graph representation is flexible and the vertices can represent any biological object. Hence a futur work is to integrate the prediction RNA-protein complex.

References

L. J. K. Aarts Emile. Local Search Algorithms. Wiley, John Wiley & Sons Ltd, 1997.

S. A. Ahmed, S. Mneimneh, and N. L. Greenbaum. A combinatorial approach for multiple rna interaction: formulations, approximations, and heuristics. In International Computing and Combinatorics Conference, pages 421–433. Springer, 2013.

M. Andronescu, V. Bereg, H. H. Hoos, and A. Condon. RNA STRAND: the RNA secondary structure and statistical analysis database. BMC bioinformatics, 9(1):340, 2008.

M. Andronescu, Z. C. Zhang, and A. Condon. Secondary structure prediction of interacting RNA molecules. Journal of molecular biology, 345(5):987–1001, 2005.

U. Benlic and J.-K. Hao. Breakout local search for the quadratic assignment problem. Applied Mathematics and Computation, 219(9):4800–4815, 2013.

E. Bindewald, K. Afonin, L. Jaeger, and B. A. Shapiro. Multistrand RNA secondary structure prediction and nanostructure design including pseudoknots. ACS nano, 5(12):9542–9551, 2011.

E. Bindewald, K. A. Afonin, M. Viard, P. Zakrevsky, T. Kim, and B. A. Shapiro. Multistrand structure prediction of nucleic acid assemblies and design of RNA switches. Nano letters, 16(3):1726–1735, 2016.

F. Glover. Tabu search, Fred Glover, Manuel Laguna, 1997.

S. Janssen and R. Giegerich. The RNA shapes studio. Bioinformatics, 31(3):423–425, 2014.

A. Legendre, E. Angel, and F. Tahi. Bi-objective integer programming for RNA secondary structure prediction with pseudoknots. BMC bioinformatics, 19(1):13, 2018.

R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA package 2.0. Algorithms for Molecular Biology, 6(1):26, 2011.

S. Mneimneh and S. A. Ahmed. Gibbs/MCMC sampling for multiple RNA interaction with sub-optimal solutions. In International Conference on Algorithms for Computational Biology, pages 78–90. Springer, 2016.

J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, and N. A. Pierce. NUPACK: analysis and design of nucleic acid systems. Journal of computational chemistry, 32(1):170–173, 2011.

Keywords: Multi RNA interaction, Structure Prediction, Secondary Structure, Maximum clique, Heuristic

Latent Tree based Inference of Ecological Network using the Poisson Log-Normal Model

Raphaëlle Momal *† [1], Stéphane Robin‡ [1], Christophe Ambroise [2]

[1] UMR MIA-Paris, INRA, AgroParis Tech, Université Paris-Saclay, 75005, Paris, FRANCE (MIA-Paris) – Institut National de la Recherche Agronomique : UMR0518, AgroParisTech, Université Paris-Saclay,Sorbonne Universités – 16 rue Claude Bernard 75231 PARIS CEDEX 05, France
[2] Laboratoire de Mathématiques et Modélisation d'Evry – CNRS : UMR8071, Université d'Evry-Val d'Essonne, Institut national de la recherche agronomique (INRA) – France

Acknowledgments
Our sincere thanks go to Corinne Vacher, who gave us access to experimental data on oak trees.

I. Introduction

In the past decade, ecological networks have become a key tool to describe interactions between species and better understand the dynamics of a whole ecosystem or anticipate its response to a given change. Such interaction networks can be inferred based on the observation of the respective abundance of each species. Metagenomics relies on Next-Generation Sequencing (NGS) technologies to evaluate the (relative) abundance of microbial species in a given medium under varying experimental conditions or across replicates. A typical metabarcoding experiment results in a vector of read counts associated with each species under study.

From a statistical perspective, network inference is usually considered in the framework of probabilistic graphical models. A huge statistical literature exists about this problem in the Gaussian case, that is when the data consists in continuous observations. These methods need to be adapted to count data.

In this work, we propose a comprehensive statistical framework for the inference of ecological networks based on metagenomic counts. To this aim, we use the Poisson log-normal (PLN) model which provides a generic description for multivariate count data. The PLN model accounts for the specificities of metagenomic data such as over-dispersion or sequencing depth heterogeneity. More importantly, the PLN model allows to correct for the effect of covariates, which is critical to avoid the detection of spurious edges in the graph.

II. Model

PLN model. The negative binomial distribution has become the reference distribution for the

*Speaker

†Corresponding author: raphaelle.momal@agroparistech.fr

‡Corresponding author: stephane.robin@agroparistech.fr

analysis of NGS read counts. This distribution is also known as the Poisson-Gamma distribution as it consists in a Poisson distribution combined with a latent Gamma layer. Unfortunately, this model does not generalizes easily to multivariate count data as no generic version of the Gamma distribution exists. The Poisson log-normal distribution is similar to the Poisson-Gamma distribution as it presents with the same over-dispersion feature, and conditional counts are Poisson distributed. However thanks to its latent log-normal layer, it easily generalizes to multivariate data via the multivariate normal distribution (Aitchison and Ho, 1989). The model can be described as follows: for each observation, a random Gaussian vector with as many dimensions as species is first drawn; each observed count is then drawn conditionally on the corresponding coordinate of the latent unobserved Gaussian vector. The dependency between the counts is therefore encoded in the covariance matrix of the latent Gaussian vector. An important feature is that, as opposed to other multivariate count distributions (Inouye & al, 2017), the correlations between species abundances can be either positive or negative, preserving the sign of the terms of the Gaussian covariance matrix.

Graphical models. A graphical model is a graphical representation of the dependency structure between a set of variables. In essence, an edge is drawn between two variables if the dependence between them does not result from the effect of the other variables. In our example, the variables are the respective species abundances and two species are connected if they are in direct interaction. One major advantage of the PLN model is that it can take advantage of methods previously developed for network inference in the Gaussian Graphical Models (GGM) framework. Our idea is to define the ecological network as the graphical model of the Gaussian latent layer of the PLN model.

Tree-based network inference. All network inference methods have to face the fact that the number of possible networks grows super-exponentially with the number of species. This makes the exhaustive exploration of the set of all possible graphs combinatorially intractable.

To circumvent this problem, we choose to model the network as a mixture on all spanning trees. This model is similar to the mixture of tree-shaped graphical models considered by (Meila & Jaakola, 2006), and allows for cycles and cliques in the network via an average over all spanning trees. This is consistent with the expectation that ecological networks are sparse, and allows us to take advantage of combinatorial results related to optimization and summation over the whole set of spanning trees.

Proposed model. Put together, the statistical model we present is a hierarchical model composed of two layers of hidden parameters:

the dependence structure, result of an average on spanning trees

the hidden Gaussian layer, conditional on the previous structure.

The observed counts are then conditional on the latent Gaussian layer.

III. Inference

Because of the presence of Gaussian latent layer, the PLN model is an incomplete data model for which the Expectation-Maximization (EM) algorithm could be considered. Unfortunately, the conditional distribution of the hidden layer given the observed data is intractable so the EM algorithm does not apply directly. However, a proxy of this distribution can be obtained using variational techniques (Wainwright & Jordan, 2008). This results in a Variational EM (VEM) that has been implemented in the 'PLNmodels' R-package available on github (https://github.com/jchiquet/PLNmodels, Chiquet, Mariadassou & Robin, 2018).

The inference of the PLN model provides an estimate for the covariance matrix of the Gaussian layer. Once this estimate is available, the inference problem takes place in the GGM context. Our method uses the set of spanning trees, which displays several interesting combinatorial features. This makes maximization (Chow & Liu, 1968) or summation (Chaiken & al, 1978) achievable in polynomial time. We compare our method with the Graphical LASSO, a commonly used network inference method.

This second layer of the model being a mixture, its inference can be carried out via an EM algorithm. Part of our contribution is to develop a new EM algorithm along which the conditional distribution of trees given the data is computed. Unlike what is usually found in the literature, the conditional probability given the data for each edge to be part of the graphical model is updated, and not fixed beforehand. Once the conditional probability of each edge is computed by averaging over all trees, the inferred graph is defined by the most probable edges.

IV. Simulation

We tested our method with several dependence structures and several densities of edges. In addition to spanning trees, we considered Erd́os structures, which are random graphs, scale-free structures which are rather sparse and clusters. The latter are very different from the other structures and should be challenging for our method. The number of vertices in the graph has been set between 10 and 30, edge probability varies between 0.025 and 0.25 and the number of observations between 20 and 100.

Considering the original graph as ground truth, our approach allows the inference of a family of nested graphs derived from the thresholding of the estimated conditional edges probabilities. The Area Under the Receiver Operating Curve (AUC) is used as a summary measure of the graph reconstruction quality. The AUC of our method was compared to that of the glasso for all settings and dependence structures. In a specific experiment, inferences are done on 40 different graphs.

Our method performs well on trees, and as expected is less efficient on the other cases but it is still comparable to or better than the glasso. In all tested settings, in terms of median of AUC our method is about 5% above glasso with trees, about 3% in the scale-free structure and only by 1% in the cluster. On the Erd́os structure the two methods perform identically. The medians of AUC are around 80% when only the number of vertices varies, however they increase significantly with the number of observations : 62% for 20 observation, and 85% for 100 with our method.

V. Illustration

The fungal Erysiphe alphitoides (EA) is the causal agent of oak powdery mildew. Jakushkin et al (2016) study its pathobiome via microbial network inference and emphasize the importance of covariates. The sampling of oak leaves microbiome was done on three different oaks with different infection status. The corresponding data table is composed of 116 samples of 94 species of fungi and bacteria of oak leaves, including the EA agent. Several covariates are available, among which the tree identifier, the distance from the leaf to the tree base, and a measure of infection.

Relative species abundances were evaluated by metabarcoding, for which it is necessary to correct for depth of coverage. Treating the later as an offset, we fitted four PLN regression models (all including offsets) on these data, including covariates one by one. They are nested and take an additional variable among those previously mentioned.

For each of the four models, we computed the conditional probabilities of edges to be part of the network. To define the threshold above which an edge is included in the network, we evaluated the overall proportion of absent edges using a multiple testing technique proposed by Storey (2002), which we cannot detail here due to space limitation.

As expected, the more covariates are included in the model, the less edges are inferred in the corresponding network, underlying the benefits of taking covariates into account. Edges removed at each step can be interpreted as spurious edges from the preceding step that were actually reflecting the effect of the included covariate. The model only adjusting for the offset contains 2630 edges, whereas the one with four covariates has 2300 edges. Between these two models all nodes lose 7 connexions on average. Regarding the pathogen EA across all models, its major role in the organization of the pathobiome is proven by its degree remaining stable at about 60 (59, 64, 63 and 60 respectively).

VI. Discussion

We provide a comprehensive statistical framework for the inference of ecological networks based on NGS read counts, which includes a formal probabilistic model and the associated estimation algorithm. Our model infers interaction networks and easily adapts to different experimental conditions by enabling the user to account for offsets and covariates.

Our final algorithm uses successively a VEM algorithm for the PLN and an EM algorithm for the inference of the tree structure. The latent layer of the PLN is first inferred using the PLNmodels package, then the EM algorithm infers the network. A technical perspective consists in building an algorithm which encompasses our EM algorithm in the M step of the VEM algorithm, within the PLNmodels package.

Finally, a challenging issue for network inference is the possibility that some species of or covariate having a strong impact on the ecosystem was not measured, resulting in spurious edges (see the illustration section). The automatic detection and estimation of such missing variable can be considered in the context of tree-shaped graphs.

Bibliography

AITCHISON, John et HO, C. H. The multivariate Poisson-log normal distribution. Biometrika, 1989, vol. 76, no 4, p. 643-653.

CHAIKEN, Seth et KLEITMAN, Daniel J. Matrix tree theorems. Journal of combinatorial theory, Series A, 1978, vol. 24, no 3, p. 377-381.

CHIQUET, Julien, MARIADASSOU, Mahendra, et ROBIN, Stéphane. Variational inference for probabilistic Poisson PCA. arXiv preprint arXiv:1703.06633, 2017.

CHOW, C. et LIU, Cong. Approximating discrete probability distributions with dependence trees. IEEE transactions on Information Theory, 1968, vol. 14, no 3, p. 462-467.

INOUYE, David I., YANG, Eunho, ALLEN, Genevera I., et al. A review of multivariate distributions for count data derived from the Poisson distribution. Wiley Interdisciplinary Reviews: Computational Statistics, 2017, vol. 9, no 3.

JAKUSCHKIN, Boris, FIEVET, Virgil, SCHWALLER, Lo´ic, et al. Microbial ecology, 2016, vol. 72, no 4, p. 870-880.

MEILă, Marina et JAAKKOLA, Tommi. Tractable Bayesian learning of tree belief networks. Statistics and Computing, 2006, vol. 16, no 1, p. 77-92.

STOREY, John D. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2002, vol. 64, no 3, p. 479-498.

WAINWRIGHT, Martin J., JORDAN, Michael I., et al. Graphical models, exponential families, and variational inference. Foundations and Trends® in Machine Learning, 2008, vol. 1, no 1–2, p. 1-305.

Keywords: Ecological networks, EM algorithm, graphical models, spanning trees, metagenomics, Poisson log normal model

On improving the approximation ratio of the r-shortest common superstring problem

Tristan Braquelaire [1], Raluca Uricaru *† [1], Mathieu Raffinot [1], Marie Gasparoux [1,2]

[1] Laboratoire Bordelais de Recherche en Informatique (LaBRI) – Univ. Bordeaux, CNRS : UMR5800 – Domaine Universitaire 351, cours de la Libération 33405 Talence Cedex, France

[2] Département dÍnformatique et de Recherche Opérationnelle (DIRO) – Université de Montréal Pavillon André-Aisenstadt CP 6128 succ Centre-Ville Montréal QC H3C 3J7, Canada

The Shortest Common Superstring problem (SCS) consists, for a set of strings, in constructing a minimum length string that contains all strings of the set as substrings. It is a crucial problem in several bioinformatics fields, among which the Next Generation Sequencing data assembly problems. While a 2.3667 [1] approximation ratio algorithm has recently been published, the general objective is now to break the conceptual lower bound barrier of 2. Here we focus on a particular instance of the SCS problem, meaning the r-SCS problem, which requires all input strings to be of the same length, r [2, 3]. In [2], Golovnev et al. proved an approximation ratio better than the general one for r $\leq$ 6. Here we extend their approach and improve their approximation ratio, which is now better than the general one for r $\leq$ 7, and less than or equal to 2 up to r = 6.

General considerations on the SCS and the r-SCS problems

An SCS greedy algorithm is known to reach good performances in practice but its guaranteed approximation ratio has only been proved to be 3.5 while conjectured 2. Improving the SCS, and in particular the r-SCS approximation ratio is interesting from both theoretical and practical reasons, in view of their numerous applications, like in bioinformatics and more precisely for the assembly problem. This problem consists of reconstructing the genome from its sequenced reads. Even if DNA reads are longer than 7 characters, our result is a step ahead in efficiently solving this problem. Gallant et al. [4] showed that the r-SCS problem stays NP-hard, except for the 2-SCS case that can be solved in polynomial time [5]. Golovnev et al. proposed an approximation ratio for the r-SCS problem, which was better than the best general approximation ratio (i.e., the one for the general SCS problem) at the time their article was published (2.4782 [6]) for r < 8. However, in the meantime, the general approximation ratio has been improved from 2.4782 to 2.3667, thus canceling their result for r = 7. In this new context, the approximation ratio proved by Golovnev et al. for the r-SCS problem remains better than the actual general one for r < 7. With our results, we extend the approach of [2] and exhibit a new approximation ratio for the r-SCS problem.

It is straightforward to see that finding a solution to the SCS problem comes to computing a Hamiltonian path of maximum weight (named H below) in the overlap graph. Indeed, H

*Speaker

†Corresponding author: raluca.uricaru@u-bordeaux.fr

would directly lead to a shortest superstring solution for the SCS problem, whose compression would be equal to the weight of H denoted by w(H). The best existing approximation algorithm [7] for computing a weighted hamiltonian path (derived from the asymmetric maximum traveling salesman path, MAX ATSP), gives a hamiltonian path whose length is at least 2/3 of the weight of the longest path, i.e., 2/3 w(H). Therefore, this generates a superstring solution that is a 2.5-approximation for the SCS problem, which is far from the actual best known approximation ratio of 2.3667.

Golovnev et al. solution for the r-SCS problem

Golovnev et al. use a (r − 1)-spectrum in order to translate the initial instance of the r-SCS problem in a 2-SCS instance, which they exactly solve with the approach described in [5]; this gives them a solution that, once translated back to the original problem, represents a good approximation of the optimal superstring for the original problem, given that the optimal is small.

Given that a k-mer is a string of length k, in Golovnev et al. the notion of k-spectrum of the input set is defined as the set of k-mers issued from the sequences of the input set.

De Bruijn graphs are largely employed in Next Generation Sequencing (NGS) data analysis, and specifically in genome assembly, as they display interesting properties like providing an intrinsic succinct representation of the data, and enabling the implementation of efficient methods for computing a superstring solution (which represents a reasonable approximation of the original genome sequence). A de Bruijn graph modeling a set of strings S is built on the k-spectrum of S as following: nodes are k-mers and oriented edges connect two k-mers if they overlap on exactly (k − 1) characters. In this work, as in [2], the de Bruijn graph is used in a particular context : given an initial set of strings of length r, a de Bruijn graph is built on the (r − 1)-spectrum corresponding to this set of strings.

The 2-SCS problem is a particular case of the r-SCS problem when r = 2, which deserves special attention since it has been shown that it is solvable in polynomial time [5], even when considering multiplicities (meaning that the strings must appear in the resulting superstring a given number of times).

Golovnev et al. translate an instance of r-SCS on a set S (composed of strings of length r) in a 2-SCS instance by computing the (r − 1)-spectrum of S and building a de Bruijn graph on this set of (r − 1)-mers. By assigning a character to each (r − 1)-mer, a string in S (originally of length r) becomes a string of length 2 in the novel alphabet. Next, they exactly solve the 2-SCS problem with an eulerian procedure (by adding minimal additional edges) [5], based on the graph illustrated in Figure 1 (obtained from the de Bruijn graph on the (r − 1)-spectrum of S). They eventually expand the resulting sequence (built on the new alphabet) by replacing each two letters (i.e., two original (r − 1)-mers) connected by an edge in the graph with their corresponding r-mer. This leads to a superstring solution for the original r-SCS problem (not necessarily optimal), we name
tau.

Figure 1 : Illustration of the resolution of the 2-SCS problem on the set {AB, BC, BD, DE,

FG, HI, JK}, by building an eulerian path (with minimal additional edges). The resulting

superstring can be obtained by traversing the eulerian path and concatenating the labels :

J→K→H→I→F→G→A→B→D→E→B→C = JKHIFGABDEBC.

A novel hierarchical approach for the r-SCS problem

In the first step of our approach we apply the same translation as in [2], meaning that from the original r-SCS problem, by using the (r – 1)-spectrum of S, we obtain a 2-SCS instance. After computing an optimal solution for the 2-SCS problem with the algorithm of [5], and then applying the reverse translation, we obtain a first, unsophisticated solution. This solution is the same as the one output by Golovnev et al., but their method stops here. In our case, we continue the initiated process by subsequently generating a set of ”contigs”, i.e., substrings of the superstring solution obtained by embedding strings from S if overlapping on exactly r – 1 characters. Note that contigs correspond to paths in the de Bruijn graph (corresponding to the 2-SCS instance), and that the length of a contig is at least r. This set of contigs is computed from the initial superstring solution by cutting the superstring up in chunks each time the connection is not due to an edge in the graph but rather to an edge added by the eulerian path resolution procedure.

For the next step we extend the notion of k-spectrum to take as input a set of contigs of possibly different lengths, but all greater than k + 1: the k-mers composing this new type of k-spectrum are the prefixes and suffixes of size k of the input contig sequences. In our case, we compute such a (r – 2)-spectrum on the set of contigs issued from the first step of the method. We then build a kind of de Bruijn graph for which the nodes come from the (r – 2)-spectrum of the contigs and for each contig sequence w, we add an oriented edge from pref(w, (r – 2)) to suff(w, (r – 2)) labeled by w. Finally, as in the first step, an eulerian path with minimal additional edges is computed on this graph, which gives a novel superstring solution for the r-SCS problem, that we call γ.

The intuition behind our algorithm is to push further the approach of Golovnev et al. by additionally taking into account the (r – 2)-overlap edges from the overlap graph built on S. However, this extension is not straightforward since (a) choices of (r – 1)-edge paths are made in the first step of our algorithm, which cannot be reconsidered in the following steps; these (r – 1)-edge paths selected in the first step prevent us to use some (r – 2) edges from the overlap graph, typically those branching inside a contig, and (b) the contigs possibly have different lengths (thus the translation into a 2-SCS instance is not straightforward).

Thorough analysis of the proposed algorithm

Golovnev et al. based their analysis on the property that the eulerian path they build on the de Bruijn graph for producing their superstring solution
tau contains all (r – 1)-overlap edges from the overlap graph, and thus at least as many as the (r – 1)-overlap edges taken in the hamiltonian path H built on the overlap graph. Our 2 steps algorithm is more difficult to analyze in the sense that our γ superstring mixes (r – 1) and (r – 2) overlaps, and, if it also contains at least as many (r – 1) overlaps as H, the number of (r – 2) overlaps can be less than that used in H, due to the fact that we do not consider all (r – 2) overlaps when building our generalized (r – 2)-spectrum.

Details of the analysis are not given here, but we are able to bound the length of γ relatively to w(H) using a bounded approximation. Indeed, as we are not able to compare exactly the number of (r – 1) and (r – 2) edges between γ and H, we consider those edges together, but counting on overlap of (r – 2) even for the (r – 1)-overlap edges. Thus, with respect to Golovnev at al., our approach further capitalizes on large overlap edges in the overlap graph but introduces a bounded approximation in the analysis, which we compare to that of Golovnev et al.

Conclusion

Our approach gives better results than that of Golovnev et al. for $5 < r < 8$. Further this limit of r = 8 we still obtain a better bound, but not better than the 2.3667 general approximation ratio. We could consider three levels instead of two, by taking into account (r – 3) edges in addition to (r – 1) and (r – 2) ones. However we observe that by extending the hierarchical approach the approximation ratio becomes worse than that of the 2-level approximation algorithm for r = 7 and r = 8.

Bibliography

K. E. Paluch. Better approximation algorithms for maximum asymmetric traveling salesman and shortest superstring. CoRR, abs/1401.3670, 2014.

A. Golovnev, A. S. Kulikov, and I. Mihajlin. Approximating shortest superstring problem using de bruijn graphs. In CPM, volume 7922 of Lecture Notes in Computer Science, pages 120–129. Springer, 2013.

J. Gallant, D. Maier, and J. Astorer. On finding minimal length superstrings. Journal of Computer and System Sciences, 20(1):50 – 58, 1980.

M. Crochemore, M. Cygan, C. S. Iliopoulos, M. Kubica, J. Radoszewski, W. Rytter, and T. Walen. Algorithms for three versions of the shortest common superstring problem. In CPM, volume 6129 of Lecture Notes in Computer Science, pages 299–309. Springer, 2010.

M. Mucha. Lyndon words and short superstrings. In SODA, pages 958–972. SIAM, 2013

H. Kaplan, M. Lewenstein, N. Shafrir, and M. Sviridenko. Approximation algorithms for asymmetric tsp by decomposing directed regular multigraphs. J. ACM, 52(4):602–626, July 2005.

Keywords: r, Shortest Common Superstring Problem, Approximation Algorithm, De Bruijn Graph, NGS

Protein domain sequence analyses using Long-Short Term Memory Recurrent Neural Networks

Tristan Bitard Feildel *† 1, Alessandra Carbone 1,2

1 Laboratoire de Biologie Computationnelle et Quantitative (LCQB) – Sorbonne Université, CNRS : UMR7238 – France

2 Institut Universitaire de France (IUF) – Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche, Ministère de l´ducation nationale, de l'Enseignement supérieur et de la Recherche – Maison des Universités 103 Boulevard Saint-Michel 75005 Paris, France

Introduction

Protein domains are sub-units of protein sequences. Due to their conservation at the sequence and structural levels they can be referred as the protein evolutionary building-blocks (Bornberg-Bauer and Alba 2013, Lees et al. 2016). Thus, a protein domain family can be defined as a group of homologous sequences found in different proteins of different organisms with similar functions.

Inside a group, proteins are likely to have similar structure despite the diversity of the protein sequences of the family. The diversity of the protein sequences inside a family reflect the different evolutionary paths taken during the molecular evolution of organisms.

*Speaker

†Corresponding author: tristan.bitard-feildel@upmc.fr

Interestingly, different positions can have different degree of conservation, some even displaying co-evolutionary pattern where the change at one position is observed along the change at another position to guarantee structural and functional conservation.
However, understanding why evolutionary changes are tolerated for some positions in the sequences and how they can be compensated is mainly not known. A better comprehension of these constrains on protein sequences is of uttermost importance for fields such as molecular evolution or protein design.

A way to statistically represent the protein sequences of a family is to model them as Hidden Markov Models (HMMs) (Eddy 1996) built from a Multiple Sequence Alignment (MSA). These models have many advantages as they permit to quickly scan large set of protein sequences for matching signatures of the families. They can also be used to generate protein sequences based on the emission/transition matrix probabilities stored inside the model. However, HMMs can be costly to compute if considering long distance relationship between columns of the MSAs and usually only the previous amino-acid is considered. Nevertheless, this trade-off lead to the creation of some of the most powerful bioinformatics tools (Eddy 1998, Remmert et al 2011) and databases for functional annotation (Oates et al 2015, Finn et al 2016).

Here, we investigate Long-Short Term Memory Recurrent Neural Network (LSTM-RNN abbrv. to LSTM in the text below) models (Hochreiter and Schmidhuber 1997) applied on protein domain family sequences. LSTM are a class of deep learning algorithms able to remember long (and short) relationships between sequential data. They proved themselves to be very powerful in Natural Language Processing tasks, speech recognition problems or video analyses due to their ability to capture long-term dependencies of sequential data.

LSTM ability to capture long term dependencies can be exploited to model biological protein sequences. We present results on different bioinformatics tasks such as sequence design or sequence classification and feedbacks regarding the implementation of deep learning frameworks. In a first task, we trained several LSTM networks, using Pfam protein families as input. These models are then used to generate protein sequences which are compared to the biological sequences and sequences emitted by the HMM of the families. The sequences are also compared based on protein structures generated by homology prediction to evaluate their biological likelihood. In a second task, LSTM networks are trained to classify the sequences of the three groups. The classifier networks are then analyzed to detect learned features by the networks used to distinguish the different groups of sequences.

Results

The 10 largest families of the Pfam database are selected for analyses. Many-to-many LSTM networks are constructed using the un-aligned sequences of each family (which correspond for most to more than 50.000 sequences). In this architecture an output vector is computed for each input of the recurrent neural network. This output vector is based on the current input character and a state vector computed from the previous inputs. The state vector is updated and passed to the computation of the output vector for the next input. Regarding amino acid, the networks are trained to predict the next amino acid (output vector) of a sequence based on the previous observed characters (the input and state vectors). Before training, protein sequences are shuffled and separated into train, validation and test sets (60%, 20%, 20% of the sequences). One-hot encoding is used to model the amino acids and different parameters regarding the size of the neural networks and the number of recurrences are tested.

After training, the LSTM networks of each family are used to generate a comparable number of protein sequences to that present in the biological protein families. These two sequence datasets are referred as *lstm* and *biological* respectively. A third sequence dataset is also generated for each family using their HMMs and referred as *hmm.*

The three datasets are compared to each other per families. Independent PCA analyses are performed on a MSA computed for each sequence datasets of each family using Mafft. We show that despite not having protein sequences in common the *lstm* and *biological* datasets display similar sequence space structure. The *hmm* sequence datasets lack of a structured space and have less explained variance by the first two components of their PCA. Interestingly, the PCAs of the *lstm* and *biological* datasets display densely populated area of similar sequences.
The sequences of the *lstm*, *biological* and *hmm* datasets of each family are used to generated protein structures.
The protein from the PDB associated to each of the Pfam families are clustered using MMseqs2 to reduce sequence and structure redundancy. The *lstm*, *biological* and *hmm* sequences are also clustered at 30% of identity.
Interestingly the number of clusters is similar between *lstm* and *biological* datasets corroborating the similar sequence space conformation between the two datasets.
Next, we analyze the quality of the sequence designed by LSTM networks and if they are closer to biological sequences than sequences generated by HMMs using these structural homology reconstruction. For each cluster a representative sequence is extracted and used as a query to search for PDB templates. PDB templates, *lstm*, *biological* and *hmm* sequences are then selected if similar hits with high coverage against the same PDB sequences are found for a sequence of each dataset.
This pipeline permits to optimize the sequence/template pairs but also to select comparable sequences from the three sequence datasets. Computing good homology models can be time consuming and is a subject by itself. The models generated in our experiment are computed using MODELLER which allows a good trade-off between speed and quality. As the the sequence/template pairs were selected based on similar sequence identity and coverage between query and the PDB template, we can focus on the observed divergence between models rather than on their intrinsic qualities. Five models per protein are computed and the DOFE score of MODELLER is used to evaluate them. A low score indicates a good reconstruction.
Interestingly, in half of the families *lstm* sequences have lower mean scores than the *biological* sequences.
Sequences from the *hmm* groups have the highest scores, i.e. the poorest quality models. This indicates that despite choosing similar best matches between the three group sequences and PDB templates, *lstm* sequences appear to be closer to *biological* sequences than *hmm* sequences .
Based on the sequence spaces and the modeling results, LSTM networks seem able to learn features which are not captured by HMMs. However, understanding how a feature, or a neuron activation, is associated to the sequence dataset context is difficult both in the interpretation of the features characterizing the datasets and in their comparison. To this aim, we built a LSTM based network classifier to distinguish *lstm* and *biological*, *lstm* and *hmm*, and *biological* and *hmm* sequence datasets and such making possible the comparisons between the activation pattern of the networks. Activated neurons to classify the sequences should highlight features which are specific to a dataset.

The classification is performed using a many-to-one LSTM architecture. In this model, the last output of the recurrent neural network, i.e. computed after passing through the whole query sequence, is used to classify the input. Thus, the context of the whole sequence is processed before doing the classification. Classifications are performed on train/validation/test subsets and different size of networks are tested with different parameters. The networks are able to easily distinguish *hmm* sequences from *biological* or *lstm* sequences (accuracy on test set higher than 95% for all families execpt 2). They also succeed to separated *biological* from *lstm* sequences but with less accuracy ($>$ 85%). The state and output vectors of each classifier were analyzed using LSTM4vis, a tool created to facilitate the interpretation of LSTM networks. They were also compared to biophysical properties of the sequences such as their hydropathy pattern, their secondary structure element assignation, the amino acid structural contacts ... Some properties can be linked to the activation of a particular neurons but the remaining neurons are hard to interpret.

Discussion

All-in-all, we show that sequences designed by LSTM networks have *biological*-like properties. The networks are able to learn the underlying sequence space of a family and generate similar biological sequences but not identical. Their capacities are confirmed using sequence-based and structure-based analyses. Access to the features learned by the network are however difficult.
Some neural activation can be linked to biological properties but most of them are hard to interpret. Finally, deep learning frameworks are relatively easy to implement. High levels API are available and benefit from the support of very active communities.
One of the major difference to overcome from ”traditional” frameworks is the extensive use of computational graphs to structure the program. The publication speed of new releases can also be intimidating, for instance Tensorflow (the Google's API) as one minor release every month.
Due to their powerful prediction and their simplicity of usage, deep learning architectures will gain in importance in bioinformatics.
New architectures and solutions to interpret their success are also very active research area and such should facilitate their adoption by the community.

References

Eddy, S. Hidden Markov Models. Current Opinion in Structural Biology 6(3):361-365 1996
Hochreiter, S. and Schmidhuber, J. Long short-term memory. Neural Comput 9(8):1735-1780 1997
Finn, R. et al. The Pfam protein families database: towards a more sustainable future. Nucleic acids research 44(D1):D279-85 2016
Oates, M. et al. The SUPERFAMILY 1.75 database in 2014: a doubling of data. Nucleic acids research 43:D227-33 2015
Eddy, E. Bioinformatics, Profile Hidden Markov Models. Bioinformatics, 14:755-763, 1998
Remmert et al HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods. 9(2):173-175 2011
Lees et al. Functional innovation from changes in protein domains and their combinations. Current Opinion in Structural Biology 38:44-52 2016
Bornberg-Bauer, E. and Alba, Mar. Dynamics and adaptive benefits of modular protein evolution. Current Opinion in Structural Biology 23(3):459-466 2013

Acknowledgement

We thank NVIDIA and the NVIDIA GPU seeding grant program for the award of a NVIDIA Titan Xp GPU.

Keywords: Protein domains, Sequence analysis, Biological sequence properties, Long Short Term Memory Networks, Deep learning

The Ocean Gene Atlas: exploring the biogeography of plankton genes online

Emilie Villar *[1], Thomas Vannier *†[1], Caroline Vernette *‡[1], Magali Lescot *§[1], Pascal Hingamp *¶[1]

[1] Institut méditerranéen dócéanologie (MIO) – Institut de Recherche pour le Développement : UMR$_D$235, *AixMarseilleUniversité* : *UM*110, *UniversitédeToulon* : *UMR*7294, *CentreNationaldelaRechercheScientifique* : *UMR*7294 – –*M.I.O.InstitutMéditerranéendÓcéanologieBâtimentMéditerranée*163*AvenuedeLuminy*13288*Marseille*, *France*

The Ocean Gene Atlas (OGA) is a web service to explore the biogeography of genes from marine planktonic organisms. It allows users to query protein or nucleotide sequences against global ocean reference gene catalogs. With just one click, the abundance and location of target sequences are visualized on world maps as well as their taxonomic distribution. Interactive results panels allow for adjusting cutoffs for alignment quality and displaying the abundances of genes in the context of environmental features (temperature, nutrients, etc.) measured at the time of sampling. The charts displayed on the OGA results page can be annotated online and downloaded as image files in vector graphics formats (SVG and PDF) suitable for publication. The ease of use enables non-bioinformaticians to explore quantitative and contextualized information on genes of interest in the global ocean ecosystem. Currently the OGA is deployed with i) the Ocean Microbial Reference Gene Catalog (OM-RGC) comprising 40 million non-redundant mostly prokaryotic gene sequences associated with both Tara Oceans and Global Ocean Sampling (GOS) gene abundances, and ii) the Marine Atlas of Tara Ocean Unigenes (MATOU) composed of > 116 million eukaryote unigenes. We reproduced the analysis of Sebastián *et al.* (*The ISME Journal* **10**:968–978, 2016) about phospholipase C abundance in low phosphate concentrations areas using OGA to verify that the web service conforms to the published results. Additional datasets will be added upon availability of further marine environmental datasets that provide the required complement of sequence assemblies, raw reads and contextual environmental parameters. Ocean Gene Atlas is a freely-available web service at: http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/.

Keywords: Ecogenomics, Environmental genomics, Visualization, Plankton, Biogeography

*Speaker

†Corresponding author: thomas.vannier@mio.osupytheas.fr

‡Corresponding author: caroline.vernette@gmail.com

§Corresponding author: magali.lescot@mio.osupytheas.fr

¶Corresponding author: pascal.hingamp@mio.osupytheas.fr

How in vitro data can contribute to in vivo toxicity prediction using Machine Learning ?

Ingrid Grenet *† 1, Jean-Paul Comet‡ 1

1 Laboratoire dÍnformatique, Signaux, et Systèmes de Sophia Antipolis (I3S) – Université Nice Sophia Antipolis, Centre National de la Recherche Scientifique – 2000, route des Lucioles - Les Algorithmes - bât. Euclide B 06900 Sophia Antipolis, France

Toxicology studies are mandatory for the marketing of chemical compounds to assess their risk of being toxic for living organisms and the environment. The most important studies, which look for absence of adverse outcomes, are conducted *in vivo* in different animal species and for different periods of time (from some days to the whole life-time of animals). These *in vivo* studies can be completed by *in vitro* ones in which several assays focus on different molecular targets and reveal compounds' bioactivity. All these studies are time, money and animal consuming which raise ethical and economical concerns leading to envision alternative solutions for toxicity evaluation : a big challenge is to assess as early as possible the potential of a chemical of being toxic, in order to avoid performing unnecessary *in vivo* studies. Indeed, this "early assessment" would help to prioritize and select compounds that are good candidates to test further in regular studies.

One possibility is to use computational methods such as machine learning (ML) to predict toxicity based on early stage data (e.g. molecular structure or *in vitro* activity). For example Quantitative Structure-Activity Relationship (QSAR) models have been developed to predict *in vitro* activity [2] and toxicity [9,6] from molecular structure. ML has also been applied to predict *in vivo* effects based on *in vitro* bioactivity data [5,7].

Nonetheless, the ideal would be to build the global chain, starting from chemical structures to predict *in vivo* toxicity through intermediate data. The goal here is to evaluate if bioactivity data obtained from *in vitro* assays can contribute to this prediction. Therefore, we propose a two-stage ML approach where the first stage, described in Section 1, is to predict *in vivo* toxicity from *in vitro* data and the second stage, described in Section 2, is to predict bioactivity based on chemical structures using QSAR methods. These two stages should then be chained up to predict *in vivo* toxicity directly from structural data for a new compound for which only the molecular structure is known (see supplemental figure). Section 3 presents results of the approach applied to a concrete example. Since we show that *in vitro* data is not sufficient to predict *in vivo* toxicity, we try in Section 4 to make this prediction either directly from the molecular structure or from the molecular structure combined with *in vitro* data.

1.Prediction of *in vivo* outcomes from *in vitro* data

This work is based on two different data sources released by the US Environmental Protection Agency : (i) Toxicity Reference Database (ToxRefDB) [4] is a public database storing results from *in vivo* toxicological studies for more than 900 chemicals. It provides either the dose at

*Speaker

†Corresponding author: grenet@i3s.unice.fr

‡Corresponding author: comet@i3s.unice.fr

which an adverse outcome effect is observed or the maximal dose for which no effect is reported. (ii) ToxCast provides results for more than 8000 chemical compounds tested in up to 800 *in vitro* high-throughput assays [1]. Results are provided as AC50 values, i.e. the concentration leading to 50% of activity of a compound in the assay. For both datasources, all data are binarised, the value 1 meaning that an adverse outcome or a bioactivity has been reported.

The first stage of the approach aims at linking *in vitro* ToxCast data to *in vivo* ToxRefDB data, based on a ML method similar to [5,7]. For any outcome of interest from ToxRefDB, ML model is built following three steps : (i) data gathering for the learning process, (ii) selection of *in vitro* assays from ToxCast which are correlated with the ToxRef outcome, and (iii) use of this subset of assays as descriptors in a ML model predicting the ToxRef outcome (see Stage1 of supplemental figure).

Data gathering for the learning process

We start by seeking the molecules having a reported value in ToxRefDB for the outcome of interest. Then, we select all assays in ToxCast in which those molecules occur. Finally, because not all the compounds have been tested in all the assays, we look for a subset of molecules and a subset of assays such that there is no missing value.

Selection of *in vitro* assays correlated with the *in vivo* outcome

The aim is to extract a subset of ToxCast assays that are related to the ToxRef outcome. Each ToxCast assay from the previous step is compared to the ToxRef outcome in a univariate way using three correlation tests (Pearson, Student and Chi-squared). An assay is considered related to the ToxRef outcome if, for at least one of the three tests, the calculated p-value is lower than a predefined cutoff. The selected *in vitro* assays then become the features describing compounds' bioactivity.

Machine learning

The dataset contains examples characterized by bioactivity descriptors (one binary value for each *in vitro* assay obtained previously) and by one binary value for the ToxRef outcome to predict. Several ML algorithms are used for the learning : Linear Discriminant Analysis (LDA), Naive Bayesian (NB), K-nearest neighbors (KNN), Support Vector Machine (SVM), Regression Trees and Random Forest (RF). Ten fold cross-validation is performed to assess the performance of the models : the initial dataset is split into a training set (90%) and a test set (10%). This process is repeated 10 times and we compute the average of three performance metrics : Sensitivity (TP/(TP+FN)), Specificity (TN/(TN+FP)) and Balanced Accuracy, named BA for short ((Sensitivity + Specificity)/2) ; where TP, TN, FN, FP denote the number of True Positives, True Negatives, False Negatives and False Positives respectively.

2.Prediction of *in vitro* bioactivity from molecular structures

The second stage of the approach aims at linking molecular structures and *in vitro* bioactivity from ToxCast. Basically, each assay correlated to the considered ToxRef outcome becomes the output to predict (see Stage2 of supplemental figure). Thus, there is one QSAR model for each assay.

Generation and selection of structural descriptors

For each *in vitro* assay from the previous stage, we use a dataset containing all the molecules that have been tested in ToxCast. All those examples are characterized by molecular descriptors automatically computed from the structure (described in Structure Data Files) and by a binary output (ToxCast assay result). About 5000 descriptors are considered : physico-chemical properties and fingerprints that are binary vectors representing the presence/absence of chemical substructures. To reduce this number, we discard descriptors whose variance is close to zero since they are not sufficiently discriminating the compounds. Also, we avoid redundancy amongst descriptors by keeping only one descriptor among all highly correlated ones (> 0.8). Finally, we perform a Fisher-test between each descriptor and the output assay, rank the descriptors according to the obtained p-value and keep the 20% best descriptors.

QSAR approach

For each *in vitro* assay, a QSAR model is built using the datasets obtained previously. The same learning algorithms and the same procedure as the first stage are applied and models' performance is assessed by the same three metrics.

3.*In vitro* assays do not contribute to long term *in vivo* toxicity **prediction**

Case study : adverse outcomes in liver of rats

Here we focus on outcomes observed in the liver of rats after 2-year toxicity studies, because on the one hand liver is an important toxicity target organ and on the other hand rat is the species for which the most data is available. We divide the reported effects into 3 outcomes [3] : hypertrophy, proliferative lesions and injury. The objective is to build and chain up the two types of ML models described previously. After having built the ML models, for each new molecule, one would use the QSAR models from Section 2 to predict *in vitro* bioactivities from the structure. Then, these predictions would become the descriptors of the second model from Section 1 to predict the *in vivo* outcome (see 3 and 4 of supplemental figure).

If the two-stage ML approach gives unsufficient performance, we could combine the two types of descriptors used (structure and bioactivity) to predict the *in vivo* outcome, as it has already been proposed by Thomas et al. [8] and Liu et al. [3].

Results of the two-stage approach

The selection of data of interest from ToxCast and ToxRef from Section 1 leads to a complete matrix of 404 compounds and 37 *in vitro* assays. The number of positive compounds (causing an adverse outcome) is 191, 116 and 133 for hypertrophy, proliferative lesions and injury respectively and the number of negative ones (not associated with any of the three chosen adverse outcomes at the doses tested) is 151. Depending on the outcome, between 25 to 30 assays are kept after univariate selection using a p-value of 0.05.

Stage 1 ML models that predict the three outcomes from the selected assays give unsatisfactory BA values between 0.5 and 0.6 depending on the algorithms. Sensitivity varies between 0.3 and 0.6 for hypertrophy and is around 0.3 for the two other outcomes. Since those results are not satisfying, we hypothezise that the *in vitro* assays used are not enough related to the outcome and does not help for the *in vivo* prediction in a ML context.

The second stage is based on unbalanced datasets (too few positive compounds compared to negative ones). Therefore, we use data augmentation techniques for balancing the training set and get better results for the 25 to 30 assays kept by the first stage. The QSAR models have an average BA, Sensitivity and Specificity of 0.65, 0.35 and 0.95 respectively. Not surprisingly, the chaining of the two stages is still unsatisfactory since the ML models predicting *in vivo* from *in vitro* does not perform well. Indeed, we obtain BA around 0.5 with Sensitivity and Specificity around 0.4 and 0.5 respectively for the entire approach and the three outcomes.

4.Direct ML from structural data to *in vivo* toxicity

Since *in vitro* assays are not enough related to *in vivo* toxicity to be used in ML, we try to predict *in vivo* toxicity either from molecular structure only or from molecular structure combined with *in vitro* data. In the first case, using the same algorithms than previously, we obtain average BA of 0.6 with Sensitivity and Specificity both varying between 0.45 and 0.75, depending on the outcome. The combination of the two types of descriptors does not improve these performances, meaning that *in vitro* bioactivity does not bring any more useful information for the ML prediction.

Conclusion

We propose here to use the early available data of chemical compounds to predict *in vivo* long term toxicity and thus help in the selection of interesting compounds. In particular, we evaluate the interest of taking into account *in vitro* bioactivity data. Thanks to a 2-stage ML approach, we chain the prediction of *in vitro* bioactivity from molecular structures with the prediction of an *in vivo* outcome from the *in vitro* bioactivity. Since models based on *in vitro* data have low performances, we could not expect satisfying results for the entire approach. Also, we obtain equal (and sometimes higher) performances for the prediction of *in vivo* toxicity directly from chemical structures. Finally, the combination of both types of data does not improve the results. Overall, these results show that the *in vitro* data used is this work cannot contribute to the prediction of long term *in vivo* toxicity. Further work may aim at using other types of early available data such as toxicogenomics or *in vivo d*ata from short term studies.

References

1.Dix et al., Toxicological Sciences, 2007

2.Hansch et al., Accounts of Chemical Research, 1993

3.Liu et al., Chemical Research in Toxicology, 2015

4.Martin et al., Toxicological Sciences, 2009

5.Martin et al., Biology of reproduction, 2011

6.Ng et al., Chemical Research in Toxicology, 2015

7.Sipes et al., Toxicological Sciences, 2011

8.Thomas et al., Toxicological Sciences, 2012

9.Zang et al., Journal of Chemical Information and Modeling, 2013

Keywords: Machine learning, toxicology, prediction, in vitro, in vivo

Causal Mediation Analysis with Multiple Mediators

Allan Jérolon * [1], Laura Baglietto , Flora Alarcon† , Vittorio Perduca‡

[1] MAP5 - Mathématiques Appliquées à Paris 5 (MAP5 - UMR 8145) – Centre National de la Recherche Scientifique : UMR8145, Institut National des Sciences Mathématiques et de leurs Interactions : UMR8145, Université Paris Descartes - Paris 5 : UMR8145 – UFR Mathématiques et Informatique, 45 rue des Saints-Pères 75270 PARIS CEDEX 06, France

Introduction: Causal mediation analysis is widely used in various domains such as biostatistics, epidemiology, psychology, legal and social sciences and public policy. The goal of such an analysis is to explain and quantify the effects of a variable on an outcome, directly and indirectly through other variables called mediators. In 2010, Imaí and collaborators introduced a general framework to define, identify and estimate these effects [1] and implemented their methods in the widely used R package *mediation* [2]. When two or more mediators are considered, current approaches consists in repeating several simple mediator analysis in parallel. This could result in an estimation bias for quantities of interest effects.In this work, contributions are threefold: First we show that conducting several simple mediator analysis in parallel result in a biased estimate of the direct effect. Then we propose a generalization of the approach by Imaí and collaborators in the case of multiple mediators which lead to unbiased estimates of direct effects. At last we implement our algorithm in R and apply it to simulate and real data.
Method: Our work is an extension of the framework of [1] in the case of multiple mediators. More precisely, we first introduce definitions of direct, indirect (mediate) and total effect, based on counterfactuals. Then we show that under proper hypothesis, these effects are non-parametric identifiable. In the case of a linear model relating the outcome with the other covariates (mediator, treatments and confounders), we show that effects are estimated very naturally using product of coefficients - type of formula as in the Linear Structural Equation modelling literature. At last we derive estimators in the case of a binary outcome when the model is either probit or logistic. All methods are implemented in R using the same quasi-Bayesian approach described in [1] and [2].

Results: Our results are for a continuous mediator and continuous or binary outcome. We validate our method on simulated data. Moreover, we show empirically that our method provides an unbiased estimate of the direct effect while estimates obtained by running in parallel simple mediator analysis are biased. At last, the proposed approach will be illustrated on a real dataset, for quantify the effects of hormone replacement treatment onto breast cancer risk through three mediators, namely dense mammographic area, non-dense area and body mass index.

References:

*Speaker

†Corresponding author: flora.alarcon@parisdescartes.fr

‡Corresponding author: vittorio.perduca@parisdescartes.fr

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological methods*, *15*(4), 309.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software,* vol. 59, issue 5.

Keywords: Mediation, Causality, Counterfactuals

Easy16S: a user-friendly Shiny interface for analysis and visualization of metagenomic data.

Cédric Midoux *† [1,2], Olivier Rué [2], Olivier Chapleur [1], Mahendra Mariadassou [2], Théodore Bouchez [1], Valentin Loux [2], Ariane Bize [1]

[1] Hydrosystems and Bioprocesses Research Unit, Irstea, France (HBAN) – Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture - IRSTEA – 1 rue Pierre-Gilles de Gennes 92761 Antony Cedex, France

[2] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France (MaIAGE) – Institut national de la recherche agronomique (INRA) : UR1404 – Bâtiment 210-233 Domaine de Vilvert 78350 Jouy en Josas Cedex, France

Microbiome data investigation has become a crucial step of recent studies of microbial diversity and dynamics, for example for environmental bioprocesses (1-3) and metabolic processes (4, 5).Studying microbial communities through next-generation sequencing henceforth often involve the analysis and interpretation of large and high-dimensional datasets. For 16S metabarcoding approaches, a two-step process is usually implemented. It consists firstly of the bioinformatics processing of nucleotide sequence files to obtain, after several operations, count and affiliation tables. Secondly, statistical analyses and visualizations are classically used to explore the data and support interpretation. Such marker gene amplicon sequencing approaches are currently affordable for most laboratories and are therefore used well beyond the community of bioinformaticians. Therefore, there is presently a high demand for user-friendly, interactive tools favoring the accessibility of data analysis to researchers with biology background.
Regarding the bioinformatics aspects, several solutions have been available for several years with command-line approach (6, 7), or through the Galaxy platform (6, 8). By contrast, tools meeting the demand for statistical analysis and visualization emerged much more recently (9-13). Shiny-phyloseq (9) is a major example of such application available to support biologists. However, from our point of view, these tools are for some aspects too complex and therefore do not exactly meet the needs of our users. To facilitate a quick and dynamic visualization of such data, we developed an interactive R-shiny interface (14) named "Easy16S". This tool is intended for biologists eager to explore their data and create figures rapidly and interactively. It is simple, easy-to-use and specifically focused on the mapping of covariates of interest.

Easy16S accepts as entry the classical BIOM output files generated by 16S metabarcoding analysis tools like FROGS (8), QIIME (6) or Mothur (7) and usually further processed with specific R packages such as phyloseq (15) (leveraging ggplot2 (16), vegan (17), ade4 (18), ape (19) and picante (20)) or mixOmics (21) for statistical analysis. Easy16S is mainly based on two R packages, shinydashboard (22) and phyloseq (15). It avoids the use of command lines while providing access to state-of-the-art methods and tools in the field. Easy16S development relied

*Speaker

†Corresponding author: cedric.midoux@irstea.fr

on a small real dataset with 18 samples, 533 taxa, 3 samples variables and a phylogenic tree.

To use Easy16S, an abundance data file in the biom format is required as primary input and a tabular metadata file (csv, tsv or excel table) as well as a phylogenetic tree (nwk format) can also be added. After formatting, the data are available as a phyloseq object. It is subsequently possible to plot various figures. Some statistical add-ons are also present. Currently, Easy16S supports summaries of count, taxonomic and sample tables; global and focused diversity histograms; hierarchical clustering of communities; α-diversity (boxplot and table); β-diversity (heatmap, network and table); rarefaction curves; phylogenetic tree browser; heatmap to visualize the count table; and various multivariate analyses (ordination, hypothesis testing, etc.). All these figures can be adjusted with imported metadata. For example, covariates of interest can be mapped to color and shape and samples can be split according to the level of another covariate. Dataset can be rarefied with a random subsampling without replacement. Tables can be filtered and ordered. Ecological distance and ordination methods can be chosen from a list for adjusting clustering, heatmap and networks. Plots and tables can be exported in both raster and vector formats.

The interface was tested on a real homemade dataset with 86 samples, 736 taxa and 19 samples covariates (unpublished dataset). Figures were plotted with low latency (few seconds), the interface was responsive and user feedbacks were very positive. This application has already been used for a user-friendly integration and visualization of bioreactor metabarcoding data (23) and many unpublished studies. It allowed end users to explore data, to plot figures with specific covariates highlighted and to perform statistical analyses.

Easy16S is currently run on an open source shiny server installed on the INRA MIGALE bioinformatics platform (http://genome.jouy.inra.fr/shiny/easy16S/). This project is currently managed in an IRSTEA GitLab repository and was written with collaborative development and the continuous addition of features to meet users' needs in mind. The next steps for Easy16S project are an LDAP user management, server resource optimization, and a full-fledged user manual.

1. Carballa M, Regueiro L, Lema JM. Microbial management of anaerobic digestion: exploiting the microbiome-functionality nexus. Curr Opin Biotechnol. 2015;33:103-11.

2. Poirier S, Bize A, Bureau C, Bouchez T, Chapleur O. Community shifts within anaerobic digestion microbiota facing phenol inhibition: Towards early warning microbial indicators? Water Res. 2016;100:296-305.

3. Koch C, Muller S, Harms H, Harnisch F. Microbiomes in bioenergy production: from analysis to management. Curr Opin Biotechnol. 2014;27:65-72.

4. Mach N, Berri M, Estelle J, Levenez F, Lemonnier G, Denis C, et al. Early-life establishment of the swine gut microbiome and impact on host phenotypes. Environ Microbiol Rep. 2015;7(3):554-69.

5. Alard J, Lehrter V, Rhimi M, Mangin I, Peucelle V, Abraham AL, et al. Beneficial metabolic effects of selected probiotics on diet-induced obesity and insulin resistance in mice are associated with improvement of dysbiotic gut microbiota. Environ Microbiol. 2016;18(5):1484-97.

6. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335-6.

7. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537-41.

8. Escudie F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, et al. FROGS: Find, Rapidly, OTUs with Galaxy Solution. Bioinformatics. 2017.

9. McMurdie PJ, Holmes S. Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking. Bioinformatics. 2015;31(2):282-3.

10. Piccolo BD, Wankhade UD, Chintapalli SV, Bhattacharyya S, Chunqiao L, Shankar K. Dynamic assessment of microbial ecology (DAME): a web app for interactive analysis and visualization of microbial sequencing data. Bioinformatics. 2018;34(6):1050-2.

11. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res. 2017;45(W1):W180-W8.

12. McIver LJ, Abu-Ali G, Franzosa EA, Schwager R, Morgan XC, Waldron L, et al. bioBakery: a meta'omic analysis environment. Bioinformatics. 2018;34(7):1235-7.

13. Zhai P, Yang L, Guo X, Wang Z, Guo J, Wang X, et al. MetaComp: comprehensive analysis software for comparative meta-omics including comparative metagenomics. BMC Bioinformatics. 2017;18(1):434.

14. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for R. 2017.

15. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 2013;8(4):e61217.

16. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York; 2009.

17. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. 2018.

18. Dray S, Dufour A-B. Theade4Package: Implementing the Duality Diagram for Ecologists. Journal of Statistical Software. 2007;22(4).

19. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289-90.

20. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, et al. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010;26(11):1463-4.

21. Rohart F, Gautier B, Singh A, Le Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017;13(11):e1005752.

22. Chang W, Ribeiro BB. shinydashboard: Create Dashboards with 'Shiny'. 2018.

23. Desmond-Le Quemener E, Rimboud M, Bridier A, Madigou C, Erable B, Bergel A, et al. Biocathodes reducing oxygen at high potential select biofilms dominated by Ectothiorhodospiraceae

populations harboring a specific association of genes. Bioresour Technol. 2016;214:55-62.

Keywords: Easy16S, metagenomics, R shiny, phyloseq, data visualization, marker gene amplicon sequencing

Let-it-bin an optimised workflow for binning metagenomic short reads from multiple samples

Quentin Letourneur * , Amine Ghozlane [1,2], Guillaume Borrel

[1] Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI) – Institut Pasteur de Paris – France
[2] Institut Pasteur – Institut Pasteur de Paris, CNRS : USR3756 – France

Metagenomic Assembled Genomes constitute a major step in the characterization of the metabolism of poorly known lineages and to predict the metabolic interactions in microbial communities.
They are obtained by calculation, involving read assembly, estimation of the abundance of contigs in multiple samples and binning. Process consisting in grouping contigs that belong to the same species based on their abundance variation among multiple samples and k-mer signal.

Given the multiplicity of programs available to perform these different steps, the assessment of the best approach is critical as well as a workflow that simplifies and optimize these calculations.

Consequently, we developed Let-it-bin that is based on nextflow and singularity to improve the scalability and the reproducibility of binning calculations. It takes raw reads as input, can do assembly with four different assemblers and binning with up to eight softwares. Resulting bins are assessed with CheckM. We applied it on a simulated dataset of 60 samples containing 39 Bacteria and one Archaea from the human gut microbiome. We focused on the capacity of binning software to separate close genomes and to manage different sequencing depth for each genome. We reconstructed 38 genomes out of 40 with the combination of CLC for the assembly and DASTool to combine results of the four best binning softwares. Their average completeness and contamination was 96.3% and 0.43%, respectively. Obtained bins are close to single genome sequencing performance.
We will also present results obtained on 190 samples of mouse gut microbiome took from five countries worldwide.

Keywords: metagenomic, binning, nextflow, singularity

*Speaker

Pixel : une solution Open Source pour l'annotation, le stockage, l'exploration et l'intégration des résultats d'analyses de données multi-omiques en biologie

Thomas Denecker * [1], William Durand [2], Julien Maupetit [2], Charles Hébert [3], Jean-Michel Camadro [4], Pierre Poulain [4], Gaëlle Lelandais [1]

[1] Institut de Biologie Intégrative de la Cellule (I2BC) – Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR9198, Université Paris-Sud - Paris 11 – Bâtiment 400, Orsay, France
[2] TailorDev (TailorDev) – TailorDev – Pépinière d'entreprises Pascalis, 10 Allée Evariste Galois 63000 Clermont-Ferrand, France
[3] Biorosetics (Biorosetics) – Biorosetics – France
[4] Institut Jacques Monod (IJM) – Université Paris Diderot - Paris 7, Centre National de la Recherche Scientifique : UMR7592 – Université Paris Diderot, Bât. Buffon, 15 rue Hélène Brion, 75205 Paris cédex 13, France

Les technologies expérimentales à haut débit, aussi appelées technologies "omiques", engendrent des quantités considérables de données (*data deluge* [1]). Les laboratoires de recherche peuvent ainsi aborder leurs problématiques scientifiques sous un regard "omique". Deux solutions s'offrent à eux. La première consiste à générer de nouvelles données expérimentales (cette possibilité est facilitée par la baisse très importante des coûts financiers associés, par exemple, au séquençage à très haut débit [2]). La deuxième solution consiste à collecter parmi les données disponibles dans les bases de données publiques (SRA ou GEO), celles qui ont un intérêt pour la question scientifique étudiée. Outre les difficultés liées à la production ou à la collecte de ces données, le défi est de les analyser puis d'intégrer les résultats obtenus. Il s'agit d'un défi majeur d'un point de vue informatique, statistique et méthodologique [3]. En effet, de nombreuses analyses pourront être réalisées sur un unique jeu de données, en utilisant de multiples outils bioinformatiques ou logiciels, avec différents ensembles de paramètres. Dans ce contexte, la problématique de l'annotation, le stockage, l'exploration et l'intégration des résultats d'analyses de données multi-omiques constitue un enjeu important en biologie.

Dans le cadre d'une modélisation *in silico* du métabolisme et de l'homéostasie du fer chez les levures pathogènes [4], la fouille de données d'une centaine d'expériences haut débit (génomique, transcriptomique et protéomique) est requise. Un travail d'amélioration de notre méthodologie de suivi des analyses et de ses résultats a été conduit avec l'aide de l'entreprise TailorDev. Ce partenariat a abouti à la création du logiciel Pixel [5] qui a pour objectif de lever les difficultés associées à la manipulation de données "omiques" massives. Il a été développé en respectant les bonnes pratiques de développement (tests unitaires et fonctionnels, intégration et déploiement continu, revue de code,...). Dans l'objectif de respecter les principes FAIR [6], Pixel a également été conçu de manière à stocker de façon structurée les informations indispensables pour reproduire à tout moment une analyse particulière. Il permet ainsi d'assurer la reproductibilité

*Speaker

des résultats. Un système hiérarchique ”d’étiquettes”, gérées dynamiquement via l’application web, permet enfin d’interroger et de filtrer facilement les résultats d’analyses présents dans le système, de les combiner et de les intégrer pour un nouveau cycle d’exploration. Un tableau de bord est aussi disponible pour suivre l’évolution des analyses (figure 1 - Supplementary data).

La principale volonté lors de la conception de ce projet était de concevoir une solution Open source. Le code source de Pixel est accessible publiquement sur GitHub (https://github.com/Candihub/pixel) et distribué sous la forme d’un conteneur Docker disponible sur DockerHub permettant le renforcement de la reproductibilité, l’accessibilité et la standardisation du projet. Ainsi, toutes les personnes intéressées par Pixel sont encouragées à le tester (il est possible d’installer une instance sur un serveur ou en local sur un ordinateur) dans le cadre de leurs activités de recherche et à partager sur Github d’éventuelles nouvelles fonctionnalités qu’elles auraient développées.

Références

The data deluge, 2012, Nature Cell Biology DOI 10.1038/ncb2558

Technology: The $1000 genome, EC Hayden. - Nature; 507 : 294 5; Mar 2014

More Is Better: Recent Progress in Multi-Omics Data Integration Methods, Huang, K Chaudhary, and LX Garmire - Front Genet, 8:84, 2017

Iron homeostasis in the pathogenic yeast Candida glabrata: lessons from multi-omics data integration, Denecker et al. - Article en cours d’écriture

Pixel: a digital lab assistant to integrate and mine biological data in multi-omics projects, Durand et al. - Article en cours d’écriture

The FAIR Guiding Principles for scientific data management and stewardship, Mark D. Wilkinson et al. ,Sci Data. 2016; 3: 160018.

Keywords: Multi, omiques, annotation, stockage, exploration, intégration, Open science

Agronomic Linked Data (AgroLD): a Knowledge-based System to Enable Integrative Biology in Agronomy

Pierre Larmande *† [1,2], Nordine El Hassouni [2,3], Aravind Venkatesan [4], Clement Jonquet [2,5], Manuel Ruiz‡ [2,3]

[1] Institut de Recherche pour le Développement (IRD) – Institut de recherche pour le développement [IRD] : UMR232, Montpellier – 911 avenue Agropolis,BP 6450134394 Montpellier cedex 5, France
[2] Institut de Biologie Computationnelle (IBC) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut National de la Recherche Agronomique, Institut National de Recherche en Informatique et en Automatique, Université de Montpellier, Centre National de la Recherche Scientifique – Building 5 - 860 rue de St Priest 34095 Montpellier, France
[3] Centre de coopération internationale en recherche agronomique pour le développement (CIRAD) – CIRAD – Av Agropolis, Montpellier, France
[4] European Bioinformatics Institute [Hinxton] (EMBL-EBI) – EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK, United Kingdom
[5] Laboratoire dÍnformatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier : UMR5506, Centre National de la Recherche Scientifique : UMR5506 – 161 rue Ada - 34095 Montpellier, France

Plant science is a multi-disciplinary scientific discipline that includes research areas such as -omics, physiology, genetics, plant breeding, systems biology and the interaction of plants with the environment to name a few. Among other things, agronomic research aims to improve crop health, production and study the environmental impact on crops. Researchers need to understand deeply the implications and interactions of the various biological processes, by linking data at different scales (e.g., genomics, proteomics and phenomics). Recent advances in high-throughput technologies have resulted in a tremendous increase in the amount of genomics or phenomics data produced in plant science. This increase, in conjunction with the heterogeneity and variability of the data, presents a major challenge to adopt an integrative research approach. We are facing an urgent need to effectively integrate and assimilate complementary datasets to understand the biological system as a whole. The Semantic Web offers technologies for the integration of heterogeneous data and its transformation into explicitly knowledge thanks to ontologies. We have developed AgroLD (the Agronomic Linked Data – www.agrold.org), a knowledge-based system that exploits the Semantic Web technology and some of the relevant standard domain ontologies, to integrate genome to phenome information on plant species widely studied by the plant science community. We present some integration results of the project, which initially focused on genomics, proteomics and phenomics. Currently, AgroLD contains hundreds millions of triples created by annotating more than 50 datasets coming from 10 data sources such as Gramene.org [1] and TropGeneDB [2] with 10 ontologies such as Gene Ontology [3] and Plant Trait Ontology [4]. Our objective is to offer a domain specific

*Speaker

†Corresponding author: pierre.larmande@ird.fr

‡Corresponding author: ruiz@cirad.fr

knowledge platform to solve complex biological and agronomical questions related to the implication of genes/proteins in, for instances, plant disease resistance or high yield traits. We expect the resolution of these questions to facilitate the formulation of new scientific hypotheses to be validated with a knowledge-oriented approach.

1. Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, et al. Gramene 2013: Comparative plant genomics resources. Nucleic Acids Res. 2014;42.

2. Hamelin C, Sempere G, Jouffe V, Ruiz M. TropGeneDB, the multi-tropical crop information system updated and extended. Nucleic Acids Res. 2013;41.

3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet [Internet]. 2000;25:25–29. Available from: http://dx.doi.org/10.1038/75556

4. Cooper L, Walls RL, Elser J, Gandolfo MA, Stevenson DW, Smith B, et al. The plant ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol. 2013;54:e1.

Keywords: Plant Molecular Biology, Data Integration Knowledge Management, Semantic Web, Linked Data, RDF, SPARQL

On the use of interaction terms in GLMs to model complex experimental designs: the example of RNA-Seq data

Hugo Varet * [1,2]

[1] Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI) – Institut Pasteur de Paris – 25-28, rue du Docteur Roux 75724 Paris Cedex 15, France
[2] Plate-forme Transcriptome et Epigénome (PF2) – Institut Pasteur de Paris – 25-28, rue du Docteur Roux 75724 Paris Cedex 15, France

Complex experimental designs are more and more used to decipher the simultaneous effect of two or more factors on the transcriptome; and several well-established tools as DESeq2 [1] or edgeR [2] use Generalized Linear Models (GLM) to handle this kind of experiment. When studying the transcriptomic response to two (or more) factors it can happen that their effects are not additive, i.e. the effect of the first one depends on the level of the second. For instance, one can imagine an experiment in which – for some genes – the drug effect depends on the strain (e.g. KO or WT). A classical and often observed approach is to test for the drug effect separately on WT and KO samples performing two independent statistical tests or analyses. Venn diagrams then allow to compare the list of genes detected differentially expressed and highlight some having a WT- or KO-specific drug response. Unfortunately, this approach can lead to many false positive and false negative genes when comparing two lists built using a threshold on the P-values. Here we illustrate this problem on a real data set and show how including and testing for the interaction term of the linear model can bring out much more relevant results.

Love M, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biology. 2014; doi:10.1186/s13059-014-0550-8

Robinson M, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009; doi:10.1093/bioinformatics/btp616

Keywords: RNA, Seq, Transcriptomics, Differential analysis, Interaction

*Speaker

Investigation of differences between long read overlappers

Pierre Marijon * [1], Rayan Chikhi , Jean Stéphane Varré

[1] Inria, Université de Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL (INRIA) – L'Institut National de Recherche en Informatique et e n Automatique (INRIA) – 40 Avenue Halley, 59650 Villeneuve-d'Ascq, France

Finding overlaps between long sequencing reads is an important step at the beginning of almost all assembly pipelines.

In 2017, Chu et la. wrote a review[1] to present and compare 5 long-read overlapping tools.Overlappers have in general higher sensitivity on synthetic data than on real data. But summary statistics do not tell us whether the same overlaps are found across all tools. In other words, do overlappers retrieve the same overlaps?

We will focus on the type of overlaps that are of interest to genome assembly tools, i.e. overlaps between suffixes-prefixes of reads (also named 'dovetails overlaps').

When looking at dovetails overlaps, we observe many discrepancies between the outputs of the overlapping tools, even between two versions of the same tool.

We observe that sensibility and precision aren't ideal metrics. In fact it would be more interesting to have different evaluation metrics that are directly relevant to the downstream application. For example, in a genome assembler it is critical to correctly detect the the longest overlaps. Another question is, can we observe specific patterns for each tool in how they detect overlaps?

A preliminary study is available at this address http://blog.pierre.marijon.fr/2018/04/13/long-reads-overlapper-compare

: http://doi.org/10.1093/bioinformatics/btw811

Keywords: long, read, overlapping, genome assembly

*Speaker

Full-length transcripts sequencing (Iso-Seq) : the Institut Curie feedback

Marc Deloger * [1,2,3], Olivier Saulnier [4,1], Juliana Pipoli Da Fonseca [5], Sonia Lameiras [5], François Prud'homme [6], Olivier Delattre [1,4], Sylvain Baulande [5], Nicolas Servant [1,2,3]

[1] Institut Curie, PSL Research University, F-75005 Paris, France – Institut Curie, PSL Research University – France

[2] INSERM, U900, F-75005 Paris, France – Inserm – France

[3] MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France – MINES ParisTech, PSL Research University – France

[4] INSERM, U830, F-75005 Paris, France – Inserm – France

[5] Institut Curie Genomics of Excellence (ICGex) Platform, Institut Curie, PSL Research University, F-75005 Paris, France – Institut Curie, PSL Research University – France

[6] Pôle Infrastructures Systèmes Réseaux, SI, DFS, DSI, Institut Curie, PSL Research University, F-75005 Paris, France – Institut Curie, PSL Research University – France

Transcripts assembly solutions from short-read RNA sequencing (RNA-seq) are actually clearly insufficient due to a lack of connectivity between reads spanning spliced junctions. In fact, transcripts assembly performance benchmark papers evaluate from 20% to 30% the recall best rate and from 20% to 60% the precision best rate. With the increasing interest in alternative splicing/polyadenylation events, proteogenomics and precision medicine, it becomes essential to increase full-length isoforms results accuracy. Thanks to long-read sequencing technologies such as Pacific Bioscience (RS II and Sequel) or Oxford Nanopore (MinION), now it is possible to directly sequences full-length transcripts without the need for assembly and imputation.
On March 2018, PacBio released the Sequel 2.1 chemistry and 5.1 software, permitting to achieve up to 20Gb per SMRT Cell. Consequently, in order to compare the consequences of this upgrade, we decided to sequence again a sample that was already sequenced before, with the previous library preparation protocols, chemistries and softwares (and with standard Illumina RNA-seq technology too).
Here we will focus on : (1) the PacBio Iso-Seq protocol enhancements and the Sequel system we have in Institut Curie's NGS core facility since July 2016 (v1.2.1 chemistry + v3.1 software), (2) the comparison between PacBio Iso-Seq and standard Illumina RNA-seq results for the same cancer sample as we know that gene families, repeated genomic regions, or alternatively spliced genes are difficult to study by transcript assembly strategies from short reads, (3) the limitations of Iso-Seq (real biological full-length ? price ? sensitivity ? error rate ? fusion detection ? quantitative ? semi-quantitative ? not quantitative at all ?), (4) the interest of using short reads in complement to overcome some long reads limitations (error-correction, quantification, sensitivity), (5) some ideas to enhance wet protocol (library normalisation, housekeeping genes depletion, capture).

*Speaker

Keywords: Long reads, IsoSeq, PacBio, RNAseq, Isoforms, Full length Transcripts

PythieVar, a collaborative tool for integrating and analyzing genetic and clinical data of rare diseases

Mélissa N'debi *† [1,2], Alexis Proust [2], Loic Foussier [4,3], Guillemette Beaudonnet [5], Anya Rothenbuhler [6], Agnès Linglart [1,6], Jacques Young [3,4], Alessia Usardi [1,6], Jérôme Bouligand [2,3], Christophe Habib [1,2], Bruno Francou [2,3]

[1] Plateforme d'Expertise Paris Sud (Peps) – AP-HP Hôpital Bicêtre (Le Kremlin-Bicêtre) – France
[2] Service de Génétique Moléculaire, Pharmacogénétique et Hormonologie (GMPH) – AP-HP Hôpital Bicêtre (Le Kremlin-Bicêtre) – France
[4] Service d'Endocrinologie et des maladies de la reproduction – AP-HP Hôpital Bicêtre (Le Kremlin-Bicêtre) – France
[3] Signalisation Hormonale, Physiopathologie Endocrinienne et Métabolique – Université Paris-Sud - Paris 11, Institut National de la Santé et de la Recherche Médicale : U1185, AP-HP Hôpital Bicêtre (Le Kremlin-Bicêtre) – France
[5] Service de Neurologie adulte – AP-HP Hôpital Bicêtre (Le Kremlin-Bicêtre) – France
[6] Service d'Endocrinologie et Diabète de l'enfant – AP-HP Hôpital Bicêtre (Le Kremlin-Bicêtre) – France

Context:

Paris-Sud University Hospitals (HUPS) gather 20 reference centers for rare disorders (CRMRs)[1]. High throughput DNA sequencing is already commonly used as a diagnostic tool by 7 of these CRMRs. This process produces a great amount of data, providing genotypic and phenotypic information, which is of particular interest in the field of rare diseases. It remains a real challenge to distinguish genetic polymorphism from causative disease variants among hundreds of identified variants. The existing databases do not fully match our needs, so at HUPS, we have developed a unique database gathering multiple phenotypic traits and genotypic data from heterogeneous and isolated patients followed by the experts of our reference centers. This database is called PythieVar and it constitutes the first step towards innovative studies focused on the correlation between rare disease patients' genotype and phenotype.

Objective:

PythieVar is designed to get a global and more accessible picture of genetic and clinical data in the field of rare disorders. The main purpose is to facilitate their diagnosis and allow the establishment of correlations between genotypes and phenotypes.

Methods:

*Speaker

†Corresponding author: ndebi.melissa@gmail.com

PythieVar was built with a Python framework called Django[2]. It enables the development of a database and its web interface at once. The PostgreSQL database is manipulated with the program and is hosted on a local server, which is accessible only through an intranet service.

Our database includes two important types of data: genetic and clinic. The first one is uploaded in the database with Variant Calling Files (VCFs), which are standardized files for annotated genetic variations. Various family of diseases, such as Peripheral Neuropathy (PN); diseases of the Calcium and Phosphorus metabolism (PhCa); Congenital Hypogonadotropic Hypogonadism (CHH); pediatric Hepatic Cholestasis (HC); Premature Ovarian Insufficiency (POI) and Disorder of Sex Development (DSD) are classified in the database. For each patient, preliminary clinical diagnosis and related data are uploaded onto the PythieVar database in a sum up format along with VCFs. Expert clinicians add disease-specific clinical data, when available, into our database, through the use of a standardized clinical information form. Employed phenotypic reference terms are obtained from HPO (Human Phenotype Ontology)[3] and the Orphanet[4] classification. Patient's data are secured onto a server and patients are pseudonymous, which means that they are only identified by identifiers and no real names are reported.

Results:

Variant Calling Files were imported into PythieVar as soon as our laboratory sequenced patients' DNA. Over the last years, more than 3000 VCFs were analyzed in our laboratory and 1753 patients included into the PythieVar database. This represents a total of 492 sequenced genes with an average of 356 variants per patient, among which approximately 35 appeared to be rare variants (*i.e.* frequency in general population below 2%) and 8 were predicted as deleterious according to various algorithms.

We have been able to collect complete clinical data for more than 200 patients, while for about 500 patients we only have preliminary clinical information.

When a variant of interest is selected, PythieVar enables the identification of other patients carrying the same variant and the analysis of their clinical presentation. Depending on the number of patients presenting the same variant, the related phenotypes and other parameters like family history, inheritance model, etc., the variant will be ranked in five classes (1. Non-causative polymorphism, 2. Likely benign variant, 3. Variant of unknown signification, 4. Likely pathogenic variant, 5. Pathogenic variant) according to ACMG (American College of Medical Genetics and Genomics)[5] classification.

PythieVar users have access to different modules. Each module allows the management of a specific query and its result.

For example, a module is dedicated to family studies. It allows us to compare variants among affected family members. This permits to identify causative rare variants within the 180 shared variants (about 10 rare) in the family. Another module sorts patients carrying variants onto a candidate gene etc.

Interest / Conclusion:

PythieVar database has now been used for over one year to gather 3-years of NGS and clinical data produced by reference centers at HUPS. The main aim of this database is to facilitate patients' diagnosis. In multidisciplinary meeting between clinicians and biologists, it is now routinely used and essential for diagnosis. In addition, Pythievar is also used for research purposes. Several studies are currently ongoing to perform correlations between genotype and phenotype and compare various sub-classes or aggregations of variants within different pathologies. PythieVar represents an innovative tool that may allow a better molecular classification and medical management for rare disorders.

Acknowledgements :

We would like to thanks Alban LERMINE, Mathieu BARTHELEMY and Vivien DESHAIES from the Plateforme de Bioinformatique de l'APHP for their support as well as Pr Anne GUIOCHON-MANTEL from the Service Génétique Moléculaire Pharmacogénétique et Hormonologie.

References:

1- Plateforme d'Expertise Paris Sud.[Website].Retrieved from http://maladiesrares-paris-sud.aphp.fr/

2- Django (Version 1.10) [Computer Software]. (2013). Retrieved from https://djangoproject.com.

3- Sebastian Ḱohler, Nicole Vasilevsky, Mark Engelstad, Erin Foster, et al. The Human Phenotype Ontology in 2017 Nucl. Acids Res. (2017) doi: 10.1093/nar/gkw1039

4- Orphadata: Free access data from Orphanet. © INSERM 1997. Available on http://www.orphadata.org. Data version (XML data version)
5- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., ... & Voelkerding, K. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine*, *17*(5), 405.

Keywords: Database, Rare Diseases, Python, Django, Targeted Exome sequencing, NGS

Systematic identification and characterization of non-coding RNAs associated with bladder cancer progression

Louis Chauvière *† [1], Tharvesh Moideen Liyakat Ali [1], Emmanuel Barillot [2], François Radvanyi [3], Claire Rougeulle [1], Céline Vallot [1,4]

[1] Centre épigénétique et destin cellulaire (EDC) – Université Paris Diderot - Paris 7, Centre National de la Recherche Scientifique : UMR7216 – Université Paris Diderot Bât. Lamarck case postale 7042 75205 Paris CEDEX 13, France

[2] Institut Curie, PSL Research University, Mines Paris Tech, Bioinformatics and Computational Systems Biology of Cancer, INSERM U900, F-75005, Paris, France – Institut Curie, PSL Research University, Mines Paris Tech, Bioinformatics and Computational Systems Biology of Cancer, INSERM U900, F-75005, Paris, France – France

[3] Compartimentation et dynamique cellulaires (CDC) – Université Pierre et Marie Curie - Paris 6, Institut Curie, Centre National de la Recherche Scientifique : UMR144 – 26 Rue dÚlm 75248 PARIS CEDEX 05, France

[4] Dynamique de línformation génétique : bases fondamentales et cancer (DIG CANCER) – Université Pierre et Marie Curie - Paris 6, Institut Curie, Centre National de la Recherche Scientifique : UMR3244 – France

Cancer development not only depends on genetic defects, but also on epigenetic changes. Indeed, DNA methylation alterations, histone modifications, chromatin remodeling or expression of non-coding RNAs can alter gene expression profiles. Numerous epigenetic modifications have been shown to be altered in cancer cells and are well documented. Long non-coding RNAs (lncRNAs) roles in cancer development are currently less known even if they may have a central impact on the control of gene expression profiles. They can activate or repress transcription, act in *trans* or in *cis* and exert their effect locally, over large domains or on entire chromosomes (as *XIST* lncRNA which accumulates and silences the entire X chromosome) [1]. Only some lncRNAs were individually described as missexpressed in cancer cells: *HOTAIR* [2], *MALAT* [3] or *ANRIL* [4] have been shown to be deregulated in cancer cells. They act as oncogenes or tumor suppressors or they regulate other oncogenes or tumor suppressors [5]. A systematic identification of lincRNAs involved in cancer progression has already been done, but assuming that lincRNAs were polyadenylated [6]. We know that some lincRNAs are not processed as mRNAs and an overview of non-polyadenylated ones is still lacking.

In the ncBlacome project, we have crossed bladder cancer progression knowledge with total RNA-sequencing of 28 bladder tumors samples and 5 controls (3 normal bladder and 2 muscle samples). The cohort encompass non-muscle and muscle invasive tumors. The technology used for the RNA-seq has the benefit of detecting transcripts that are not polyadenylated, allowing long intergenic non-coding RNAs (lincRNAs) transcriptome reconstruction.

[*] Speaker

[†] Corresponding author: louis.chauviere@univ-paris-diderot.fr

We have first reconstructed the bladder cancer lincRNA transcriptome using StringTie and selected *de novo* transcripts that have lincRNAs characteristics : intergenic entities with a length of over 200 nucleotides. StringTie reconstruction was then compared to the Gencode_v24 annotation. Entities that overlapped annotated transcripts were discarded. Finally, 1723 *de novo* lincRNAs that are expressed in bladder cancer or normal bladder were detected. In addition, 7680 Gencode_v24 annotated lincRNAs were used in this study.

Among these transcripts, we have selected the 1000 lincRNAs with the most variable expression across samples and then performed an unsupervised clustering using an Euclidean distance matrix and a Ward clustering. Results finally showed that we could divide the basal cancer samples in two groups (named group A and B). This partition has never been described studying mRNA or annotated lncRNA transcriptome [7]. In the group A of basal samples, 10 lincRNAs are clearly over-expressed in basal bladder tumors and specifically in group A. Four of them have never been annotated and are poorly expressed in normal bladder tissues. They can be considered as *de novo* lincRNAs in basal bladder tumors. CGH arrays were previously performed on all the samples determining Copy Number Variations (CNVs). These data showed that lincRNAs are not localized in regions with DNA copy number changes. For a part of them, expression is generally highly correlated with their neighbor genes. This suggests that they could act locally as regulators of their neighbors genes. Results were validated using the TCGA bladder cancer cohort (410 bladder cancer samples). With our clustering method, we can find two groups of basal bladder cancers whereas only one group was previously found using the same cohort [7]. Four lincRNAs are specifically over-expressed in both group A of ncBlacome and TCGA cohorts. The 10 lincRNAs candidates will be functionally validated in bladder cancer cell lines and their role in basal bladder cancer progression will be determined.

1. Kugel, Jennifer F., and James A. Goodrich. 2012. "Non-Coding RNAs: Key Regulators of Mammalian Transcription." Trends in Biochemical Sciences 37 (4). Elsevier Ltd: 144–51. doi:10.1016/j.tibs.2011.12.003.

2. Gupta, Rajnish A., Nilay Shah, Kevin C. Wang, Jeewon Kim, Hugo M. Horlings, David J. Wong, Miao Chih Tsai, et al. 2010. "Long Non-Coding RNA HOTAIR Reprograms Chromatin State to Promote Cancer Metastasis." Nature 464 (7291). Nature Publishing Group: 1071–76. doi:10.1038/nature08975.

3. Ji, Ping, Sven Diederichs, Wenbing Wang, Sebastian B́oing, Ralf Metzger, Paul M. Schneider, Nicola Tidow, et al. 2003. "MALAT-1, a Novel Noncoding RNA, and Thymosin β4 Predict Metastasis and Survival in Early-Stage Non-Small Cell Lung Cancer." Oncogene 22 (39): 8031–41. doi:10.1038/sj.onc.1206928.

4. Gil, Jesús, and Gordon Peters. 2006. "Regulation of the INK4b-ARF-INK4a Tumour Suppressor Locus: All for One or One for All." Nature Reviews Molecular Cell Biology 7 (9): 667–77. doi:10.1038/nrm1987.

5. Arun, Gayatri, Sarah D. Diermeier, and David L. Spector. 2018. "Therapeutic Targeting of Long Non-Coding RNAs in Cancer." Trends in Molecular Medicine 24 (3). Elsevier Ltd: 257–77. doi:10.1016/j.molmed.2018.01.001.

6. Iyer, Matthew K., Yashar S. Niknafs, Rohit Malik, Udit Singhal, Anirban Sahu, Yasuyuki Hosono, Terrence R. Barrette, et al. 2015. "The Landscape of Long Noncoding RNAs in the Human Transcriptome." Nature Genetics 47 (3): 199–208. doi:10.1038/ng.3192.
7. Robertson, A. Gordon, Jaegil Kim, Hikmat Al-Ahmadie, Joaquim Bellmunt, Guangwu Guo, Andrew D. Cherniack, Toshinori Hinoue, et al. 2017. "Comprehensive Molecular Characteriza-

tion of Muscle-Invasive Bladder Cancer." Cell 171 (3): 540–556.e25. doi:10.1016/j.cell.2017.09.007.

Keywords: long non coding RNA, bladder cancer, total RNA, seq

The Role of User-Centred Design When Revisiting a Scientific Web Application : Redesign of iPPI-DB, a database for modulators of Protein-Protein Interactions

Rachel Torchet * [1]

[1] Institut Pasteur [Paris] – Institut Pasteur de Paris – 25-28, rue du docteur Roux, 75724 Paris cedex 15, France

Authors : Rachel Torchet1,, Alexandra Moine-Franel1,2,3,, Hélène Borges1,2,3, Olivia Doppelt-Azeroual1, Fabien Mareuil1, Hervé Ménager1, and Olivier Sperandio1,2,3*

1 C3BI, Institut Pasteur 28 rue du Dr Roux, 7015 Paris

2 Structural Bioinformatics Unit, Institut Pasteur 28 rue du Dr Roux, 7015 Paris

3 CNRS UMR 3528, Institut Pasteur 28 rue du Dr Roux, 7015 Paris

*To whom correspondence should be addressed.

These authors have equally contributed

Scientific Context

Pharmaceutical innovation is still impaired by the paucity of clinically testable targets and by the fact that only a few are successfully exploited in each therapeutic area. This stands in sharp contrast with the number and diversity of roles of Protein-Protein Interactions (PPIs) in cells. Indeed, with about 130,000 binary PPIs and possibly more just in humans, the development of drugs targeting these systems represents a significant step toward expanding the druggable genome and a possible leverage on the pharmacological modulation of disease-associated cellular pathways.

The key of success in finding chemical probes for PPI targets is twofold. Firstly, which PPI target should be selected for such purpose as we now know that not all PPIs are suited for pharmacological modulation? Secondly, once the PPI target has been selected, which chemotypes should be used to hit this target as it is well documented that commercial chemical libraries are poorly suited for such an endeavor?

iPPI-DB, for inhibitors of Protein-Protein Interaction DataBase, a web application first released in 2012 [1], which stores physicochemical and pharmacological data about PPI modu-

*Speaker

lators and their targets. Users can query the database using either pharmacological criteria, or by the chemical similarity [2] with respect to a user-defined query compound. Currently the database is manually curated from the scientific literature and contains 1756 non-peptide inhibitors (iPPI) across 18 families of Protein-Protein Interactions. The chemical structures, as well as the physicochemical and the pharmacological profiles of these compounds and their targets, are in the present version manually extracted from the literature, computed and retrieved using various manual steps. This rather tedious procedure seriously hinders the updates of the database.

As we stand in the post genomic era, numerous studies are carried out cumulatively providing gigantic amounts of publicly relevant data. Manually maintaining databases that are at the crossroads of different scientific fields is in this context nearly impossible, if it is not assisted with the proper technology.

To facilitate the query of the data, as well as the growth of the database, we decided to completely reinvent iPPI-DB. The redesign of the database and its web application enable, while preserving the essence of manual curation, both a more intuitive exploration of the data, and largely automate the entry of new data. Applying User-Centered Design (UCD) methodologies, we designed a contributor interface which can be used by experts of the community to quickly add new relevant data, without requiring any technical knowledge about the database itself. These contributions are then to be reviewed by the core curators of the database, whose validation of these entries makes them publicly available through the web interface.

2. A User-Centered Design Approach

For this project we applied a User-Centered Design (UCD) approach and methodology to redesign the iPPI-DB web application. The main goal of this redesign is to focus on the needs of the user, ensuring that the end product is fit for the purpose, increasing the number of entries in the database and ease the query process.

User research [3] is the process of figuring out how people interpret and use products and services, and it helped us define three different roles for users, Core Curators, External Contributors and Common Users, with their respective motivations.

We used different user experience methodologies, such as Six Up and One Up [4], to design mockups and prototypes for the different pages. Although this approach has been previously applied in some bioinformatics projects [5], it remains largely unusual [6]. Furthermore, our experience shows that the process itself is easy to set up with biologists and engineers and is highly effective. For instance, to design the query mode interface, scientists and engineers were asked to draw different versions of the same page which highlight new priorities and concerns. As a result, within two meetings we created up a consensus prototype to implement and to test. We adopted a similar approach for the maintenance interface, designing it with biologists and going through an iterative process to facilitate its use; the process may have been longer, given the novelty and complexity of this part of this application which was not redesigned but rather completely new.

Throughout the project, our overarching concern was the ergonomics of the new version. To validate this aspect, we adopted an iterative approach, interleaving successive series of design, tests and implementation steps, involving users in each iteration. This approach, although required an important involvement from the users during the project, has been extremely beneficial, as it allowed us to build a constructive dialog between scientists and the development team, and quickly validate or ask for corrections in the software.

3. A community-based web application

The redesign of the web application is driven by a community-based purpose, for the query interface that allows simultaneous combinations of many chemical and pharmacological criteria, and for a dedicated maintenance interface to enter new data in the database. This constitutes an important improvement over the previously complex and largely manual process.

The revisited query interface now allows users to combine multiple filters to build complex queries, then share the query as a URL with collaborators, and download corresponding data. Query results can be displayed as thumbnails, as a list of cards, or as a table, all sortable.

The interface for maintenance has been designed with the aim of facilitating as much as possible the contributions, even from experts who are not familiar with the technical aspects of this database. Each contribution is based on the description of the content of a publication or a patent. Through a wizard-based web interface, users provide in a step-by-step process, the architecture of the PPI complex(es), the chemical compounds tested for modulation, and the various assays in which those compounds were tested.

4. Technical implementation

The new version of iPPI-DB interface has been implemented using the Django framework. Different web services have been plugged to iPPI-DB in order to reduce the risks of errors and facilitate contributions, mostly using the BioServices package.

Since some steps of the iPPI-DB environnement require chemoinformatics tools to calculate physico-chemical properties of the different compounds, Java programs have been developed using the JChem libraries from Chemaxon and implemented into two workflows in Galaxy.

5. Conclusion

User-Centered Design helps improving user experience and adoption through the creation of well-fitted and understandable user interfaces. These aspects are extremely important to promote community contributions to this database, UCD is a real beneficial to design and redesign scientific web applications and software.

References

Labbé,C.M. et al. (2013) iPPI-DB: a manually curated and interactive database of small non-peptide inhibitors of protein–protein interactions. Drug Discov. Today, 18, 958–968.

Labbé,C.M. et al. (2016) iPPI-DB: an online database of modulators of protein–protein interactions. Nucleic Acids Res., 44, D542–D547.

Goodman E., Kuniavsky M., Moed A., 2012, Observing the user experience : a practitioner's guide to user research, 2nd edition, Elsevier.

Bowles C., BoxJ., 2010, Undercover User Experience Design, 1st edition, Berkeley, New Riders.

Néron B., Ménager H., Maufrais C., Joly N., Maupetit J, Letort S., Carrere S., Tuffery P., Letondal C., 2009, Mobyle: a new full web bioinformatics framework, Bioinformatics, 25, 3005–3011.

Pavelin K, Cham JA, de Matos P, et al. : Bioinformatics meets user-centred design: a perspective. PLoS Comput Biol. 2012;8(7):e1002554. 10.1371/journal.pcbi.1002554

Keywords: User, Centred Design, UX Design, Web application, Structural Biology

HOMARD : High throughput Optical MApping of Replicating DNA

Nikita Menezes * [1], Francesco De Carli [2], Wahiba Berrabah [3], Valérie Barbe [3], Auguste Genovesio [4], Olivier Hyrien [5]

[1] Institut de Biologie de l'ENS (Paris - Ulm) (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris, CNRS : UMR8197, Inserm, Université de recherche Paris Sciences Lettres (PSL) – 46, rue d'Ulm 75005 Paris, France

[2] Institut de Biologie de l'ENS (Paris - Ulm) (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris, CNRS : UMR8197, Inserm – 46, rue d'Ulm 75005 Paris, France

[3] Genoscope - Centre national de séquençage [Evry] (GENOSCOPE) – Commissariat à l'énergie atomique et aux énergies alternatives : DSV/IG – 2, rue Gaston Crémieux CP5706 91057 EVRY Cedex, France

[4] Institut de Biologie de l'ENS (Paris - Ulm) (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris, CNRS : UMR8197, Inserm : U1024 – 46, rue d'Ulm 75005 Paris, France

[5] Institut de biologie de l´cole normale supérieure (IBENS) – École normale supérieure - Paris, Institut National de la Santé et de la Recherche Médicale : U1024, Centre National de la Recherche Scientifique : UMR8197 – 46, Rue d'Ulm75005 Paris, France

DNA replication is a vital process that ensures an accurate conveyance of the genetic information to the daughter cells. In eukaryotic organisms, genome replication is carried out by using multiple start sites, also known as replication origins. During the S phase of the cell cycle, these origins are fired stochastically giving rise to bidirectional replication forks that propagate along the genome until converging replication forks merge; this action is called the termination [1-3].
In metazoans, the mapping of replication remains challenging. For instance, genome wide mapping of human replication origins performed using sequencing of Okazaki fragments (Oka-seq) [4] , initiation sites (Ini-seq) [5], isolated small nascent strands (SNS-seq) [6] or replication bubbles (Bubble-seq) [7], only modestly agree [2-4]. One possible explanation of this inconsistency, is that these existing genome wide approaches to mapping DNA replication use large cell populations that smooth out variability between chromosomal copies. Thus, to get a better understanding of DNA replication and to uncover the cell-to-cell variability, the development of single molecule techniques is fundamental. DNA combing is a widespread technique used to map DNA replication at a single molecule level. Unfortunately, it has a very low throughput and tends to give fainted signal in addition to uneven linearity of the DNA molecules, making the automated detection and mapping of the DNA fragments an arduous task. Therefore, single molecule techniques tend to be refractory to automation, forestalling genome-wide analysis.

To overcome these impediments, we repurposed an optical DNA mapping device based on microfluidics, the Bionano Genomics Irys system [8], for High-throughput Optical MApping of Replicating DNA (HOMARD). Relying on the same labeling strategy as OMAR [9] (Optical Mapping of DNA replication), our methodology labels fluorescently DNA replication tracks and

[*] Speaker

nicking endonuclease sites (barcode) using two different fluorescent nucleotides (dUTP), in addition to a YOYO-1 intercalator DNA fiber staining. These labelled DNA fibers are driven by electrophoresis into the nanochannels arrays of the Irys chip where they are linearized and imaged automatically. We typically collect, for a single run, over 34 000 images divided into a fixed number of "scans" and more than 63 000 Mbp of DNA. The advantages of such technology, at the raw images level, are the homogeneity of DNA molecules linearization, the lower background and the signal quality improvements. In addition, optical mapping algorithms enables an automated alignment of the DNA molecules on a reference genome thanks to the barcode signal. Thus, high throughput data collection and automation of the analysis is achievable. Our new open source tools, that required the adaptation of the provided proprietary software, empower us to map the intensity profiles extracted after having two essential preprocessing steps on the raw images. First, the illumination bias is corrected by subtracting the median image computed on each scan of a given run. Thus, we manage to increase the signal to noise ratio. Second, the chromatic shift correction, also called image registration, consist in aligning the replication signal (red) and the barcode (green) onto the DNA molecules (blue). In doing so, signal omission or erroneous collection from the neighboring DNA molecules are avoided. We can now simultaneously visualize the intensity profiles of all mapped DNA molecules, check the optical mapping performed and, in particular, see where the replication tracks are located genome-wide at a single molecule level.
We demonstrate the robustness of our approach by providing an ultra-high coverage (23,311 x) replication map of bacteriophage DNA in Xenopus egg extracts and the potential of the Irys system for DNA replication and other functional genomic studies apart from its standard use meaning genome assembly and structural variation analysis.

1. Hyrien, O. & al (2013). From Simple Bacterial and Archaeal Replicons to Replication N/U-Domains. Journal of Molecular Biology, 425, 4673–4689.

2. Hyrien, O. (2015). Peaks cloaked in the mist: The landscape of mammalian replication origins. Journal of Cell Biology, 208(2), 147–160.

3. Dewar, J. M., & Walter, J. C. (2017). Mechanisms of DNA replication termination. Nature Publishing Group, 18(8), 507–516.

4. Petryk & al (2015). Replication landscape of the human genome. Nature Communications.

5. Langley, A. R. & al. (2016). Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). Nucleic Acids Research, 44(21), 10230–10247.

6. Vassilev,L. and Johnson,E.M. (1989) Mapping initiation sites of DNA replication in vivo using polymerase chain reaction amplification of nascent strand segments. Nucleic Acids Res., 17, 7693–7705.

7. Mesner,L.D. & al. (2013) Bubble-seq analysis of the human genome reveals distinct chromatin-

mediated
mechanisms for regulating early- and late-firing origins. Genome Res., 23, 1774–1788.

8. De Carli, F. & al (2015). Single-molecule, antibody-free fluorescent visualisation of replication tracts along
barcoded DNA molecules. The International Journal of Developmental Biology, 304(May), 1–9.

9. Hastie, A. R & al. (2013). Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate De Novo Sequence Assembly of the Complex Aegilops tauschii Genome. PLoS ONE, 8(2).

Keywords: DNA replication, High, throughput, Image analysis, Optical mapping

Calcul haute performance pour l'exploration de génomes obtenus par séquençage en cellule unique avec des données métagenomiques et métatranscriptomiques de Tara Oceans.

Artem Kourlaiev * [1], Yoann Seeleuthner [2], Léo D'agata [1], Marc Wessner [1], Benjamin Noel [1], Olivier Jaillon [2], Patrick Wincker [2], Jean-Marc Aury [1]

[1] Commissariat à l'Energie Atomique (CEA), Genoscope, Institut de Biologie François-Jacob – CEA Evry 2 rue Gaston Crémieux 91006 Evry cedex – F-91057 Evry, France, France
[2] Génomique métabolique (UMR 8030) – Commissariat à l'énergie atomique et aux énergies alternatives : DRF/IG, Université d'Évry-Val-d'Essonne, Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR8030 – GENOSCOPE, 2 rue Gaston Crémieux 91057 Evry Cedex, France

Les séquenceurs haut-débit actuels génèrent une quantité de données qui croît de manière exponentielle. Pour assurer la production et l'analyse de ces données de manière efficace, les équipes du Genoscope ont recours au calcul haute performance en utilisant les supercalculateurs du Centre de Calcul Recherche et Technologie (CCRT), composante du complexe de calcul scientifique du CEA. Situé à Bruyères-le-Châtel (Essonne) dans les locaux du Très Grand Centre de Calcul du CEA (TGCC). Depuis 2013, le CCRT est la plateforme centrale d'hébergement et de traitement des données de génomiques, issues du projet national " France Génomique ".
Le changement d'échelle des analyses à effectuer, notamment dans le cadre du projet Tara *Oceans*, nécessite la mobilisation de plusieurs milliers de cœurs en parallèle et pose de nouveaux challenges, non seulement pour appliquer des traitements bioinformatiques sur ces importantes masses de données mais également pour visualiser et interagir avec les résultats des analyses de manière efficace et globale.

Le projet Tara *Oceans* (2009-2013) a pour objectif d'étudier les écosystèmes planctoniques marins (des virus aux métazoaires). Des échantillons d'eau ont été prélevés puis filtrés sur des stations couvrant l'ensemble des océans. Le séquençage de l'ADN et de l'ARN présents dans ces échantillons a été effectué au Genoscope et a généré plusieurs milliers d'échantillons. Ce projet constitue le plus grand effort de séquençage jamais réalisé pour des organismes marins.

La métagenomique est une méthode permettant d'étudier l'ensemble des génomes des populations de micro-organismes d'un écosystème donné à partir d'un échantillon environnemental. La métatranscriptomique permet de se focaliser sur les gènes et leur expression. Ce sont les deux approches choisies pour étudier la biodiversité dans le cadre du projet Tara *Oceans*.

*Speaker

Le séquençage en cellule unique permet de séquencer l'ADN extrait à partir d'une cellule unique, prélevée dans un échantillon environnemental. Cette méthode est utilisée pour l'étude précise des génomes de protistes marins non cultivables en laboratoire et présents dans les prélèvements d'eau de l'expédition Tara *Oceans*.

Combinant les données 'meta-omique' avec les données générées par du séquençage en cellule unique, il possible d'explorer les génomes d'organismes non cultivables. Ainsi, dans un premier temps les données metagenomiques ont été utilisées pour faire une biogéographie. Cette biogéographie a été réalisée sur l'ensemble des organismes séquencés en cellules uniques. Dans un second temps, les données metatranscriptomiques des échantillons ont été utilisés pour produire une annotation structurale de ces génomes. Ceci apporte un catalogue de gènes jusque là inconnus et offre des nouvelles voies d'analyses.

L'alignement de séquences est couramment utilisé dans ce genre d'analyses mais la volumétrie des données mobilisées, plus de 800 milliards de lectures, et le nombre de génomes, plus d'une centaine, a nécessité la mise en place de méthodes de parallèlisation massive sur les supercalculateurs du CCRT-TGCC.

Dans un premier temps, l'ensemble des outils nécessaires aux analyses ont été identifiés, installés et testés sur le centre de calcul. Les tests effectués sur les données préliminaires ont permis d'optimiser les installations, les traitements appliqués sur les données et d'estimer les ressources nécessaires en termes d'heures de calculs, d'espace de stockage et de réseau.

Dans un second temps, les traitements ont été automatisés en utilisant plusieurs techniques de parallélisations. La parallélisation à mémoire partagée (multithreading et flux) a été utilisée pour les traitements ne nécessitant qu'une seule machine. La parallélisation à mémoire distribuée a été utilisée pour des traitements nécessitant une parallélisation sur plusieurs machines. Cette automatisation tient compte de l'architecture du centre de calcul et du volume de données mobilisé. Ainsi, elle permet de traiter de manière efficace l'ensemble des données avec des temps de restitution courts. A titre d'exemple, un alignement de 1 000 échantillons métagenomique de 400 millions de lectures chacun sur un assemblage de 50Mb prend moins de 10 heures en parallèlisant sur 1 400 coeurs.

Enfin, dans un troisième temps, les données nécessaires aux analyses postérieurs sont archivés sur un système de stockage hiérarchique (mise sur bande automatique) ayant une capacité de plusieurs Péta-octets. Les données nécessaires à la visualisation interactive des résultats sont ensuite générées et copiées au Genoscope sur des serveurs dédiées.
Grâce à une conception massivement parallèle et à l'utilisation des moyens de calculs du TGCC-CCRT, nos workflows fonctionnent sur des infrastructures de plusieurs dizaines de milliers de coeurs, permettant de diminuer le temps de restitution. A titre d'exemple, nous pouvons aligner l'ensemble des données métagenomique et métatranscriptomique du projet Tara *Oceans* sur un génome en quelques heures. La méthode mise en place est extensible, et pourra être utilisée pour d'autres projets de ce type.

Keywords: metagenomique, metatranscriptomique, single, cell, hpc, high performance computing, calcul à haute performance, big data, data science

Florilege: an integrative database using text mining and ontologies

Estelle Chaix* [1], Sandra Dérozier †‡ [1], Louise Deléger [1], Hélène Falentin [2], Jean-Baptiste Bohuon [1], Mouhamadou Ba [1], Robert Bossy [1], Delphine Sicard [3], Valentin Loux [1], Claire Nédellec [1]

[1] Unité Mathématiques et Informatique Appliquées du Génome à l'Environnement – MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France – France
[2] Science et Technologie du Lait et de lÓeuf (STLO) – Institut National de la Recherche Agronomique : UMR1253, Agrocampus Ouest – 65, rue de Saint Brieuc 35042 Rennes, France
[3] Sciences Pour l'Oenologie (SPO) – Institut National de la Recherche Agronomique : UR1083, Institut de Recherche pour le Développement, Université de Montpellier, Université Montpellier 1, Institut national d'études supérieures agronomiques de Montpellier – France

In Life Science, huge amount of critical information is published in many databases and scientific publications. This is also the case in the microbial biodiversity field of such as that studied in food microbiology.
Textual information is under-exploited because it is expressed in natural language, and so unstructured. The main sources are scientific articles, but also free text fields of databases. The objective of this work is to provide a unified access to the information of these diverse sources, being textual or not.

This work focuses on the extraction of relevant information on microorganism of food products, with an emphasis on positive flora. Indeed, food fermentation and biopreservation processes involve the use of various species and strains of bacteria and yeast. These strains are responsible for the targeted characteristics of the food products that are sanitary, organoleptic (aroma and texture) and healthy properties [Marco *et al.*, 2017].

Previous work has shown that the relationships between microbes, their living place (food as Habitat), and their phenotypic properties are information of interest for biologists [Chaix *et al.*, 2017]. For information extraction purpose, we designed a corpus composed of scientific article references from PubMed bibliographic database and text fields of Biological Resource Center catalogues, e.g. Inra CIRM (Centre International de Ressources Microbiennes, http://www.inra.fr/cirm), DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen) [S´ohngen *et al.*, 2015] and major genetic databases, e.g GenBank, BioSample, as part of the Florilege project.

The text-mining process behind information extraction has been set up by Inra using Alvis tools [Bossy *et al.*, 2015] in the OpenMinTeD environment (http://openminted.eu/).The treatment applied to the corpus is the same whatever the source of the textual information: (i)

*Corresponding author: estelle.chaix@inra.fr
†Speaker
‡Corresponding author: sandra.derozier@inra.fr

entities detection of relevant parts of text, that are words or word groups, are identified and assigned to a type, "taxon", "habitat" or "phenotype" ; (ii) normalization assigns a category of the relevant knowledge resource to the identified entities), (iii) and finally, relationship extraction links these entities together. The result is stored in the Florilege database.

Microbial taxa are assigned categories of the NCBI reference taxonomy, e.g. 1639 is the Listeria monocytogenes ID in the NCBI taxonomy. Food products (microbial Habitat) and Phenotypes are automatically categorized according to the OntoBiotope ontology [Nédellec *et al.*, 2017]. For example, various text extracts, "traditional soft Churpi cheese of Yak milk", "Reblochon" or "soft white Hispanic-style cheese" are assigned the same habitat reference category that is "soft cheese", according to the OntoBiotope ontology. Such formalization of unstructured text is the key point of integration of heterogeneous data in the Florilege database.

The Florilege application displays a unique set of structured information (822,006 links between 81,740 "taxa" and 2,342 "habitats") on microbial food flora, publicly accessible at http://migale.jouy.inra.fr/Flo It offers numerous cross-functional avenues of exploitation in different fields like food security, ecology, and human health. We believe Florilege will be a highly valuable tool to (i) assess phenotypic biodiversity of food microbes (ii) assign biochemical functions to each strain/species from fermented or biopreserved food products (iii) help into the development of innovative food products in particular those that involve fermentation or biopreservation processes.

References

Marco M.L., Heeney D., Binda S., Cifelli C.J., Cotter P.D., Foligné B., G´anzle M., Kort R., Pasin G., Pihlanto A., Smid E.J. and Hutkins R. (2017). **Health benefits of fermented foods: microbiota and beyond**. Current Opinion in Biotechnology, 44:94–102.

Chaix E., Aubin S., Deléger L., & Nédellec C. (2017, July). **Text-mining needs of the food microbiology research community**. In 2017 EFITA WCCA Congress.

S´ohngen C., Podstawka A., Bunk B., Gleim D., Vetcininova A., Reimer L. C., ... & Overmann J. (2015). **Bac Dive–The Bacterial Diversity Metadatabase in 2016**. Nucleic acids research, 44 (D1), D581-D585.

Bossy R., Golik W., Ratkovic Z., Valsamou D., Bessieres P., & Nédellec C. (2015). **Overview of the gene regulation network and the bacteria biotope tasks in BioNLP'13 shared task**. BMC bioinformatics, 16(10), S1.
Nédellec C., Bossy R., Chaix E., Deléger L. (2017) **Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity**. Proceedings of the 4th International Microbial Diversity Conference.

Keywords: text mining, database, ontologies

A data science approach for exploring differential expression profiles of genes in transcriptomic studies – Application to the understanding of ageing in obese and lean rats in the FIGHT-HF project.

Emmanuel Bresso * [1], Claire Lacomblez [1], Anne Pizard [2], Patrick Rossignol [2], Faiez Zannad [2], Malika Smaïl-Tabbone [1], Marie-Dominique Devignes [1,3]

[1] Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA) – Centre National de la Recherche Scientifique : UMR7503, Université de Lorraine, Institut National de Recherche en Informatique et en Automatique – Campus Scientifique BP 239 54506 Vandoeuvre-lès-Nancy Cedex, France

[2] CIC-Nancy – Institut Lorrain du Coeur et des Vaisseaux Louis Mathieu [Nancy], Institut National de la Santé et de la Recherche Médicale : CIC1433 – 4 rue du Morvan - Bâtiment Louis Mathieu - 54500 Vandoeuvre-les-Nancy Cedex, France

[3] ESPRI-Biobase [CHRU Nancy] (Unité fonctionnelle de la plateforme d'aide à la recherche clinique) – Centre Hospitalier Régional Universitaire de Nancy – Hôpital Central - 29, avenue du Maréchal de Lattre de Tassigny - CO 60034 - 54035 Nancy Cedex, France

Introduction

Transcriptomic studies are known to produce huge amounts of information about differentially expressed genes in various situations. The expression levels measured for several thousands of genes in contrasting pairs of situations (different tissues or organs, different physiological state, different age, etc.) allow to calculate fold-change ratios and false discovery rates. Today, data science should be able to derive valuable knowledge units from all these data, by extracting relevant lists of differentially expressed genes and interpreting them. However, in many cases, only a small proportion of all transcriptomic results is finally exploited.

Here we revisit the concept of differential expression profile (DEP) and we propose a combined database - network approach to extract relevant lists of differentially expressed genes based on DEP sharing and to interpret these lists using heterogeneous graph settings before visualisation.

Methods

Differential Expression Profile definition.

A transcriptomic study is usually performed on various biological samples derived from tissues or organs under certain conditions. In most cases, expression profiles are defined in a non-supervised manner by clustering the genes on the basis of their expression values across all situations in a study. In other cases, the study involves contrasting situations that lead to the calculation of fold-change (FC) ratio with false discovery rates (FDR) for each gene. Such a

*Speaker

setting allows to define differential expression profiles (DEPs).

Let a transcriptomic study be represented by a set of contrastsC_i, i∈{1...n}, defined as ordered pairs of situations, and by the FC ratios and FDR values obtained for each gene g_kin each contrast C_i. Four discrete statuses and two modalities have been defined to represent the behaviour of a given gene in a given contrast. The two modalities correspond to stringent (FC2 and FDR < 0.05) or not stringent (any FC and FDR < 0.05) definitions of differential expression. Under stringent modality, status Str1 means differential expression, status Str2 and Str3 are for up (FC$>$ 2) and down (FC$<$ -2) regulation respectively and status Str4means no differential expression. Under non stringent modality, similar statuses can be defined and are prefixed by NStr.

Formally, a DEP is described by a definition and an extension. The definition of a DEP can be represented as a set of pairs(C_i,status), i∈{1...n}and status∈{Str1...Str4,NStr1...NStr4}describing the differential expression status of a gene across a set of contrasts C_i. The extension of a DEP is the set of genes g_kthat match the definition. It should be noted that a DEP can also be defined from the differential expression statuses of a given gene across a set of contrasts and then used to select other genes matching the DEP definition.

DEP Relational Data Model.

A relational database named DEPdb was built to store and query transcriptomic results w.r.t. DEPs. The data model covers the contrast definitions between pairs of situations, the transcripts involved and their corresponding genes along with human orthologs when necessary, the FC ratios and associated FDR values, calculated for each transcript and each contrast.

A DEP query interface was implemented on DEPdb to retrieve lists of genes matching a given DEP definition or retrieve a DEP definition for a given gene.

Network-based interpretation of lists of genes.

Various tools exist already to help biologists interpret a list of genes in the light of pathways and interaction networks available in various integrated and curated public sources. Hence, complex graphs are produced showing the multiple interactions existing between the genes of interest and their interactants. Network science should help end-users to interpret such complex networks.

We present here an analysis strategy based on the notion of " heterogeneous graph " in which different types of nodes are interconnected by various types of relationships. Such heterogeneous graph is the basis of the EdgeBox, a " graph knowledge box " constructed by the EdgeLeap company, using the Neo4J graph database system and available public resources [1]. The EdgeBox currently contains seven types of nodes: protein/gene, disease, pathway, drug, metabolite, gene ontology term and miRNA, and fourteen types of relationships between these nodes (five monopartite, such as protein-protein interactions, and nine bipartite ones). The 2017 version for the FIGHT-HF project concerns human proteins/genes and counts about 211,000 nodes and almost 22 million of relationships.

To interpret a DEP extension, i.e. a list of genes corresponding to a given DEP definition, in the light of the EdgeBox, a first approach consists in querying the EdgeBox for pathways that interconnect at least two genes/proteins of the list. The query result can be completed with the genes/proteins that interact with at least two genes/proteins of the list. When the gene list is short (less than 10 genes), the resulting graph can be interpreted manually and reveals at a glance which pathways or interactions possibly explain why genes share a given DEP definition. This network-query tool has been implemented onto DEPdb connected to the graph knowledge box. It displays the heterogeneous graph using the Cytoscape program for further analyses.

When the gene list grows, the resulting graph becomes extremely complex and impossible to interpret manually. We propose to filter the nodes for enhancing user interpretation. Enrichment analyses are first carried out on the list of genes to select a small number of nodes corresponding to the top10 of significantly ($p < 0.001$) enriched pathways and biological process GO terms. In parallel, it reveals useful to select subgroups of gene/protein nodes of interest based on their neighbourhood in the EdgeBox. Such selection may be driven by user knowledge. After node reduction, a collection of smaller heterogeneous subgraphs is produced that usually become tractable.

Our combined database-network approach can function iteratively. Starting from a first DEP definition involving a given set of contrasts, the database will return a list of genes that is subsequently displayed as a heterogeneous network. In some cases, the user will select from this network a gene of interest and will query the database to retrieve its DEP across a different set of contrasts. The database will then return on demand all other genes matching this second DEP definition. The new returned extension can then in turn be analysed as a heterogeneous graph, etc.

Results and discussion: Differentially expressed genes upon ageing in heart and kidney of obese and lean SHHF rats

Description of the study

A transcriptomic study aimed at characterizing simultaneously metabolic syndrome and cardiac, vascular and renal phenotypes in ageing lean and obese SHHF (Spontaneously Hypertensive Heart Failure) rats has been described previously [2]. Obesity is induced in SHHF rats by homozygous inactivation of the leptin receptor gene. Rats have been monitored during 11 months (from the age of 1.5 to 12.5 months). In the frame of the FIGHT-HF project, transcriptomic results of this study are newly investigated using our coupled database/network approach.

Definition of two DEPs and extraction of corresponding gene lists

Transcriptomic results have been stored in the DEPdb database as described above. To illustrate our approach we focus on a simple comparison between ageing in obese versus lean rats, in both heart and kidney tissues. We therefore use four different contrasts from our database: heart samples from " old versus young " lean and obese rats (contrasts C_13 and C_14 respectively in DEPdb), and kidney samples from " old versus young " lean and obese rats (contrasts C_21 and C_22 respectively in DEPdb).

We query DEPdb consecutively with two DEP definitions, namely:

DEP_lean={(C_13,Str1),(C_14,Str4),(C_21,Str1),(C_22,Str4)},

and DEP_obese={(C_13,Str4),(C_14,Str1),(C_21,Str4),(C_22,Str1)}. In other words, DEP_lean will retrieve from DEPdb all genes differentially expressed (stringent defintion) upon aging in lean ((C_13,Str1) and (C_21,Str1)) but not in obese ((C_14,Str4) and (C_22,Str4)) rats, in heart and kidney samples respectively, and DEP_obese will retrieve from DEPdb all genes differentially expressed upon aging in obese ((C_14,Str1) and (C_22,Str1)) but not in lean ((C_13,Str4) and (C_21,Str4)) rats, in heart and kidney samples respectively. Note that the expression statuses in all other contrasts are not considered here. We retrieved 7 and 55 genes for DEP_lean andDEP_obese definitions, respectively. We also determined with a third appropriate DEP definition that 11 genes are differentially expressed upon aging in both lean and obese rats in heart and kidney samples.

Interpretation with heterogeneous graphs

We subsequently analysed the two extensions of the DEP_lean and DEP_obese profiles in the light of our network knowledge box. Please note that these gene lists first need to be converted to their human orthologs. The DEP_lean extension includes 7 genes and is short enough to be directly analyzed in DEPdb connected with the EdgeBox. We observe that all gene nodes are interconnected together through pathways and GO terms related to cell cycle and cell proliferation, metabolism of proteins and DNA, apoptosis, rhythmic processes, signal transduction and immune system.

Because of its complexity, the DEP_obese extension of 55 genes underwent the filtering process described above before heterogeneous graph retrieval from the EdgeBox. Interestingly the resulting network clearly reveals two modules of interconnected genes, with 4 genes forming bridges between the two modules. One of these modules involves 22 genes interconnected with the neutrophil degranulation pathway and several GO terms related to inflammation, whereas the other one involves 15 genes interconnected with two pathways : extracellular matrix organization and elastic fiber formation, and with GO terms related to either extracellular matrix organisation or epoxygenase P450 pathway that is involved in anti-inflammatory response.

In summary the comparison of DEP_lean and DEP_obese extensions reveals that a group of 7 genes involved in particular in cell cycle and cell proliferation is dysregulated in heart and kidney upon ageing in lean rats but not in obese rats. On the contrary, genes that are dysregulated in heart and kidney upon ageing in obese rats but not in lean rats fall into two modules, one related to inflammatory response and the other one related to extracellular matrix organisation. This analysis brings new precise molecular support to the general statement that ageing can have quite different outcomes in lean or obese individuals. The four proteins bridging these two modules (CD44, INTBD2, ANXA2 and CP2E1) can be further investigated to better understand the complementarity between the two groups of biological processes dysregulated upon ageing in the heart and kidney of obese rats. Interestingly, context analysis of the two lists of genes in a more classical framework such as CPDB returned similar results [3].

Conclusion and Perspectives

A formal definition of DEPs has been proposed for transcriptomic studies involving contrasting situations. A coupled database/network approach has been designed and implemented to explore any desired DEP from a study, by extracting the corresponding list of genes and interpreting them using a network knowledge box and heterogeneous graphs. The ageing case study shows that this approach provides useful support for manual interpretation of gene lists and can lead to new hypotheses generation. Obviously, a large number of different DEPs remain to be tested in the same way from the SHHF study.
We are currently searching to design and implement automatic filters to reduce heterogeneous graph complexity in a knowledge-based manner for assisting user interpretation of heterogeneous networks.

Reference:

Pinet F, CuvelliezvM , Kelder T, Amouyel P, Radonjic M, Bauters C. (2017)
Integrative network analysis reveals time-dependent molecular events underlying left ventricular remodeling in post-myocardial infarction patients.
Biochimica et Biophysica Acta (BBA) – Molecular Basis of Disease 1863(6): 1445-53.

Youcef G, Olivier A, Nicot N, Muller A, Deng C, Labat C, Fay R, Rodriguez-Guéant RM, Leroy C, Jaisser F, Zannad F, Lacolley P, Vallar L, Pizard A. (2016)
Preventive and chronic mineralocorticoid receptor antagonism is highly beneficial in obese SHHF rats.
Br J Pharmacol. 173(11):1805-19.

Kamburov A, Wierling C, Lehrach H, Herwig R. (2009)
ConsensusPathDB–a database for integrating human functional interaction networks.
Nucleic Acids Research 37(Database issue):D623-D628.

Keywords: differential expression profiles, heterogeneous graphs, complex network, transcriptomics

ALFA: compute and display mapped reads distribution by genomic categories and biotypes

Mathieu Bahin * [1], Benoit Noel [1], Charles Bernard [2], Valentine Murigneux [3], Leila Bastianelli [3], Herve Le Hir [3], Alice Lebreton [1], Auguste Genovesio† [1]

[1] Institut de Biologie de l'ENS (Paris - Ulm) (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris, CNRS : UMR8197, Inserm : U1024 – 46, rue d'Ulm 75005 Paris, France

[2] Institut de Biologie de l'ENS (Paris - Ulm) (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris, CNRS : UMR8197, Inserm – 46, rue d'Ulm 75005 Paris, France

[3] Institut de Biologie de l'ENS (Paris - Ulm) (IBENS) – Ecole Normale Supérieure de Paris - ENS Paris, CNRS : UMR8197, Inserm – 46, rue d'Ulm 75005 Paris, France

The last ten years have witnessed the rise of a myriad of applications that take advantage of Next-Generation Sequencing (NGS) technologies. In the vast majority of cases, whatever the species, whatever the sequencing technique, the first analysis step of this type of data consists of a quality control of the reads while the second step consists of a mapping of those reads to a reference genome. However, the subsequent steps are often very specific to the type of NGS experiment.

With this work, we aim at introducing a third systematic step after mapping which would be common to any NGS experiment. This step consists in producing a global overview of the distributions of the mapped reads across genomic categories (stop codon, 5'-UTR, CDS, intergenic, etc.) and biotypes (protein coding, miRNA, ncRNA, etc.) at nucleotide resolution. Our approach turns out to be very useful for a broad range of NGS applications we are dealing with, as it brings a sort of post-mapping quality control and a first global functional insight. In any case, it adds information to the usual mapped/unmapped read count and other post-mapping statistics.

A few tools providing this type of information have been proposed in the literature for specific NGS applications. For instance, Homer or CEAS, dedicated to ChIP-seq data, count detected peaks found in each of a predefined set of categories. However, as those tools cannot conveniently deal with mapped reads, their application to other sequencing techniques is precluded. In fact, to the best of our knowledge, there is no available ready-made tool that proposes such a quantitative overview at a nucleotide precision. Furthermore, using directly the mapped reads allows us to propose a framework working for any species and whatever the sequencing technique.

The tool we propose works in two steps. First, a provided genome annotation file (GTF format) is processed to generate an index. Each nucleotide of the genome is annotated according

*Speaker

†Corresponding author: auguste.genovesio@ens.fr

to a standard priority definition between features. Then the program computes the nucleotide fraction mapped to each predefined feature in one or more BAM files. By default, the program outputs a raw count and a normalized count plots for the categories and respectively for the biotypes. The normalization is achieved according to the relative importance of a given category or biotype in the genome in order to provide a view in term of enrichment.

We will show results obtained by the proposed tool on various types of NGS experiments such as: 1) RIP-Seq data on Saccharomyces cerevisiae samples to quantify whether the IP and Input reads are equally represented in the 3'-UTR region of the genes, 2) MeRIP-Seq data on Arabidopsis Thaliana samples to identify the type of RNA preferentially methylated and 3) Ribosome Profiling on human and mouse data to discover at an early step of the analysis that some low quality samples should be discarded.

We will show results obtained by the proposed tool on various types of NGS experiments: CLIP-Seqs data on Mus Musculus samples, BS-Seq data on Arabidopsis thaliana samples, Ribo-Seq data on Homo Sapiens and Mus Musculus samples, ChIP-Seq data on Caenorhabditis elegans and RNA-Seq data on a Saccharomyces cerevisiae sample. All those examples highlight different ALFA usages: quality control, contamination detection, first biological insight, etc.

Overall, we present a versatile, open source and freely available tool that is of a potential wide interest for the bioinformatics community.

Keywords: NGS, Quality control, Post mapping, Tool, Generic

Indexing de Bruijn graph with minimizers

Antoine Limasset * [1], Rayan Chikhi , Rob Patro , Fatemeh Fatemeh

[1] Université Libre de Bruxelles (ULB) – Avenue Franklin Roosevelt 50, 1050 Bruxelles, Belgium

Indexing de Bruijn graph with minimizers
Antoine Limasset, Fatemeh Almodaresi, Rayan Chikhi and Rob Patro
April 2018

A simple but fundamental need when dealing with genomic sequences is to be able to index very large
sets of fixed length words (k-mers) and to associate information to these sequences (origin, abundance,
strand, score etc. . .). As trivial as this need may seem, computationally challenging instances are extremely
common in Metagenomic, pangenomic and even for the study of single large genomes, where sets of dozens
or hundreds billions k-mers need to be processed. We propose a novel data structure able to both test the
membership and associate information to the k-mers of a De Bruijn graph in a very efficient and exact way.
We wrote a proof of concept dubbed Blight available at https://github.com/Malfoy/Blight to assess the
performances of our proposed scheme. We were able to index all the k-mers of a human genome with less
than 8GB of memory (≈ 24 bits per k-mer). Our proposed index is also designed to provide extremely fast
and parallel operations, able to perform billions queries in minutes.

1
Context

The de Bruijn graph structure is increasingly used as an efficient mean to represent a set of k-mers of
interest. Several previous studies focused on the representation of a set of k-mer (Gosamer (1), Minia (2),
DBGFM (3)). If those structure are extremely memory efficient (A modified version of Minia (4) achieved
the rate of 8.58 bit per k-mer on a human dataset) they do not allow to associate information to k-mers.
Recently, the usage of efficient minimal perfect hash function (MPHF) library allowed the indexation of
billions of keys with moderate resources (5). But such functions are not able to recognize alien keys that
were not in the indexed set. If such a key is queried, the MPHF may return the position of an

*Speaker

indexed key hence producing a false positive (FP) error. The trivial solution would be to associate to each k-mer, in addition to its associated value, the k-mer sequence itself. This way an alien key could be recognized. But such structure require $2 \star k$ bits per k-mer which can be extremely expensive, especially for large k. In order to cope with this problem SRC (6) proposed the use of a binary fingerprint, in order to keep the FP rate as low as possible while presenting a low memory overhead. The fingerprint mechanism lead to a n bits per k-mer overhead for a false positive rate around 1/2 n which can guarantee a very low amount of errors with a low memory cost. Another proposition made by Pufferfish (7) is to propose an exact structure also based on this MPHF in order to index specifically the k-mer of a de Bruijn graph. Their idea to handle the alien keys in a memory efficient way is to associate to each k-mer its position in the indexed De Bruijn graph. This led to a memory efficient and fast to query structure able to index a human genome with 12 GB (which represent approximatively 4 bytes per k-mer) while being two time faster than FM-index based tools as BWA (8).

2 Methods

In the Pufferfish scheme, the main memory usage come from the encoding of the positions of the k-mers in the graph as each position cost O(ln(Genome size)) to encode. We propose to improve this scheme by working on subgraphs in order to reduce the memory amount required to encode such positions. For this we will take advantages to the fact that overlapping k-mers tend to share minimizers (9) and that we can 1represent a set of n overlapping k-mers sharing a minimizer with a super-k-mer of length n+k –1. This super k-mer representation were notably used by KMC2 (10) in order to highly reduce the disk usage of external memory k-mer counting operations. The idea to improve the Pufferfish scheme come in two steps. First we will split the k-mers of our de Bruijn graph according to their minimizers, and encode them as super-k-mers. This way we have to deal with order of magnitude smaller sequences sets that we will call buckets. For example with a minimizer size of 12 on a human genome graph counting 2.5 billions k-mers, the largest bucket presented only 121,452 nucleotides. Those buckets are henceforth order of magnitude smaller and the amount of bits necessary to encode a position into them will be drastically reduced: log2(2.5 $\star$ 10 9) = 31 where log2(1.2$\star$10 5) = 17. In a second part we will encode the k-mers positions into their respective buckets, as we can know a k-mer’s minimizer directly from its sequence. This lead to several improvements :

- The amount of bit used to encode a position is highly reduced
- The data locality of the query is greatly improved, as each minimizer use its own small structure that can fit in cache, several successive queries will therefore be treated without cache miss
- The construction of the index may be done in parallel
- The membership queries can be highly optimized by using the graph structure

We implemented this method in a C++ library without any dependencies in order to be easily usable for most users. The code is open-source and available at https://github.com/Malfoy/Blight.

3 Results

We were able to index all k-mers of a human genome with less than 8GB of memory (less than 24 bits per k-mer) and the index can be built in less than one hour on a 20 cores cluster. The query of the whole dataset against itself were done within 5 minutes on the same cluster.

References

T. C. Conway and A. J. Bromage, ”Succinct data structures for assembling large genomes,” Bioinformatics, vol. 27, no. 4, pp. 479–486, 2011.

R. Chikhi and G. Rizk, ”Space-efficient and exact de bruijn graph representation based on a bloom filter,” Algorithms for Molecular Biology, vol. 8, no. 1, p. 22, 2013.

R. Chikhi, A. Limasset, S. Jackman, J. T. Simpson, and P. Medvedev, ”On the representation of de bruijn graphs,” Journal of Computational Biology, vol. 22, no. 5, pp. 336–352, 2015.

K. Salikhov, G. Sacomoto, and G. Kucherov, ”Using cascading bloom filters to improve the memory usage for de brujin graphs,” Algorithms for Molecular Biology, vol. 9, no. 1, p. 2, 2014.

A. Limasset, G. Rizk, R. Chikhi, and P. Peterlongo, ”Fast and scalable minimal perfect hashing for massive key sets,” arXiv preprint arXiv:1702.03154, 2017.

C. Marchet, L. Lecompte, A. Limasset, L. Bittner, and P. Peterlongo, ”A resource-frugal probabilistic dictionary and applications in bioinformatics,” arXiv preprint arXiv:1703.00667, 2017.

F. Almodaresi, H. Sarkar, and R. Patro, ”A space and time-efficient index for the compacted colored de bruijn graph,” bioRxiv, p. 191874, 2017.

H. Li, ”Aligning sequence reads, clone sequences and assembly contigs with bwa-mem,” arXiv preprint arXiv:1303.3997, 2013.

M. Roberts, W. Hayes, B. R. Hunt, S. M. Mount, and J. A. Yorke, ”Reducing storage requirements for biological sequence comparison,” Bioinformatics, vol. 20, no. 18, pp. 3363–3369, 2004.

S. Deorowicz, M. Kokot, S. Grabowski, and A. Debudaj-Grabysz, ”Kmc 2: fast and resource-frugal

k-mer counting," Bioinformatics, vol. 31, no. 10, pp. 1569–1576, 2015.
2

Keywords: Indexing, graph, minimizers, genomic, metagenomic

TOGGLe, a framework for simple and fast building or reproducible parallel workflows for large scale analyses

Christine Tranchan-Dubreuil [1], Sébastien Ravel [2], Cecile Monat [3], Gautier Sarah [4], Abdoulaye Diallo [5,6,7], Laura Helou , Alexis Dereeper [8], Ndomassi Tando [9], Julie Orjuela-Bouniol [10], François Sabot *† [1,11]

[1] Diversité, adaptation, développement des plantes (DIADE) – Institut de recherche pour le développement [IRD] : UR232, Université Montpellier II - Sciences et techniques – Centre IRD de Montpellier 911 av Agropolis BP 604501 34394 Montpellier cedex 5, France

[2] Biologie et génétique des interactions plantes-parasites pour la protection intégrée (BGPI) – Institut national de la recherche agronomique (INRA) : UR0385, Centre de coopération internationale en recherche agronomique pour le développement [CIRAD] : UMR54 – Campus International de Baillarguet - TA 41 / K - 34398 Montpellier Cedex 05, France

[3] IPK – Germany

[4] Amélioration Génétique et Adaptation des Plantes Méditerranéeennes et Tropicales (AGAP) – Montpellier SupAgro, Institut national de la recherche agronomique (INRA) : UMR1334, CIRAD-BIOS – TA A-108/03-Avenue Agropolis, 34398 Montpellier Cedex 5, France

[5] Etudiant en première année du parcours Bioinformatique, Connaissances, Données (BCD) du master Sciences et Numérique pour la Santé (SNS) – Université Montpellier II - Sciences et Techniques du Languedoc – France

[6] UMR DIADE IRD/UM/CIRAD,IRD - France Sud – Christine Tranchant-Dubreuil, James Tregear – 911 Avenue Agropolis, BP 64501, 34394 Montpellier Cedex 5, France

[7] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université Montpellier II - Sciences et techniques, Alban Mancheron – 860 rue de St Priest, CC 05016, 34095 Montpellier Cedex 5, France, France

[8] CIRAD UMR AGAP (AGAP) – Institut national de la recherche agronomique (INRA) : UMR1334 – TA A-108/03-Avenue Agropolis, 34398 Montpellier Cedex 5, France

[9] Institut de Recherche pour le Développement (IRD [France-Sud]) – Institut de recherche pour le développement [IRD] – 911 avenue Agropolis,BP 6450134394 Montpellier cedex 5, France

[10] ADNid (ADNid) – Aucune – ADNid company 830 Avenue du Campus Agropolis Baillarguet 34980 Montferriez sur Lez France, France

[11] South Green (SG) – Institut de recherche pour le développement [IRD], Institut national de la recherche agronomique (INRA), CIRAD, Bioversity – France

Introduction

Advances in Next-Generation Sequencing (NGS) technologies have provided a cost-effective approach to unravel many biological questions, and revolutionized our understanding of Biology. Nowadays, any laboratory can be involved in large-scale sequencing projects, delivering astonishing volumes of sequence data. Although NGS are powerful technologies, they shifted the paradigm from data acquisition to data management, storage and in fine biological analyses [7]. This intensifies the need for robust and easy-to-use pipelines to perform high-performance

*Speaker

†Corresponding author: francois.sabot@ird.fr

automated analyses [3]. However, available pipelines depend on the sequencing method used to generate raw data and on the type of analyses to perform (variant calling, GWAS, differential gene expression,...) [2,10,6,1].

TOGGLe [9] offers a robust and scalable bioinformatic framework for a wide range of sequence-based applications. TOGGLe is highly flexible on the data type, working on sequencing raw data (from Illumina, ONT, PacificBiosciences...), as well as on various other formats (e.g. SAM, BED, VCF). Carrying out analyses does not require any programming skills, but only basic Linux ones. With TOGGLe, scientists can create robust, customizable, reproducible and complex pipelines, through an user-friendly approach, specifying the software versions as well as parameters.

Implementation and Tools Input data formats

Input data format can be either FASTA, FASTQ (paired-end, single-end and mate-pair; first-, second- and third-generation), SAM, BAM, BED, GFF or VCF, either plain or compressed (i.e. gzip). Sample IDs are automatically managed by TOGGLe using the file read name, and no dedicated nomenclature or external sample declaration is needed. For pair-end/mate-pair FASTQ data, no specific name or directory organization is required for pairs to be recognized as such.

Running a TOGGLe pipeline

TOGGLe workflows can be launch from start-to-end with a single command-line, with three mandatory arguments:

the input directory containing files to analyze, the output directory that will contain results generated by TOGGLe,the configuration file.

According to the workflow (e.g. a reads mapping step upon a reference), a transcriptome or genome reference sequence, an annotation file or a key file (for demultiplexing) may be also provided.

Configuration file

This basic text file is composed of different parts allowing to build the workflow, to provide software parameters, to compress or remove intermediate data, and to set up a scheduler if needed [9].

Building Workflow

Steps composing the pipeline (e.g. aligning reads upon a reference genome, calling variants, assembling) and their relative order are defined after the $order tag. Each line consists of the step number followed by an equal symbol then by the software's name (e.g. 1=FastQC). If the step number is lower than 1000, the analysis step is carried out for each sample separately, while the step is performed as a global analysis with the results of all the samples for a value higher or equal to 1000 [9].

Providing software parameters and external tools usage

The syntax for setting software parameters is identical to that used by each software using the command line. If no software parameter is provided, the default ones are used. TOGGLe will handle itself the input and output files as well as the references. Users can also use any software not included in TOGGLe with the generic tag followed by the command-line.

Compressing or removing intermediate data

As analyses generates a large amount of data, we included the possibility to gzip compress or to remove some intermediate data ($compress and $cleaner tags).

Setting up jobs scheduler

When analyzing on high performance computing (HPC) systems, TOGGLe runs seamlessly with either LSF, MPRUN, SLURM or SGE jobs schedulers. In addition, users can provide specific environment variables to be transferred to the scheduler (such as the paths or modules to be loaded). Finally, node data transfer is automatically managed by TOGGLe when requested by user.

Workflows Management

The core of TOGGLe is the toggleGenerator.pl script which (i) reads the configuration file, (ii) generates pipeline scripts, and then (iii) executes them as parallel or global analyses [9]. Basically, toggleGenerator.pl acts as a Make-like tool, compiling blocks of code (themselves allowing the execution of the different tools) to create the requested pipeline. It allows the developers to easily add any new tool without having to modify the main code.

Platforms, Installation and Customization

TOGGLe currently runs on any recent GNU/Linux system (Debian Lenny and more, Ubuntu 12.04 and more, and CentOS 6 and more have been tested). TOGGLe was developed to be straightforward to install in several ways : manually (git clone) or through a bash script. A unique file (localConfig.pm) needs to be filled at installation to ensure the integration of the whole software list (path and version; installed separately). However, the whole set of integrated tools is not required to run TOGGLe: one can use it only with SAMtools [5,4] for instance, and does not need to install the other tools. More detailed information on the different installation procedures can be found at the TOGGLe website (http://toggle.southgreen.fr)

Analyses and post-analysis tools integrated in TOGGLe

Developed in Perl, TOGGLe incorporates more than 120 functions from 40 state-of-the-art open-source tools and home-made ones[9].

A large array of tools are ready to use with TOGGLe for various type of analyses: input data QC control, cleaning FASTQ, mapping, post-mapping treatment, SNP calling and filtration, structural variation analyses, assembly (genome and transcriptome). Post-analysis tools are available for population genetics, genomic duplication, phylogeny or transcriptomics [9].

A tool targeting both biologists and bioinformaticians Ease of use

TOGGLe drastically simplifies NGS analyses (such as SNP calling, differential expression for RNA, in silico assembly) for biologists. Workflows can be easily set up in a few minutes through a unique configuration file, and can be executed through a short command line. In addition, TOGGLe offers access to all parameters without restrictions (or name change) proposed by each software. Finally, users can provide any reference files, without any additional step to add them. Numerous pre-defined validated configuration files are available on our website (http://toggle.southgreen.fr/) for various type of analyses.

Ease of development and evolution: Simpler is Clever

TOGGLe is designed as a set of separated modules/functions and blocks of code, simplifying code integration and evolution. Each module is written either to run bioinformatic softwares or to ensure functionalities for a specific purpose (such as checking input file formats). The block files are composed of codes implementing a single function at a time. These blocks are then concatenated together following the user pipeline specifications by toggleGenerator.pl, to provide a dedicated script pipeline.

This code modularity as well as testing and development processes adopted in TOGGLe prevents the regressions and bugs, facilitating maintenance in a collaborative environment.

A robust bioinformatics framework

As TOGGLe was developed initially for performing data-intensive bioinformatics analyses, our main aim was to build a robust workflow framework without sacrificing the simplicity of use and the ease of development.

Pipeline and data sanity controls

Numerous automatic controls are carried out at different levels as transparent actions : validation of the workflow structure defined by the user (checking if the output file format by step n is supported by the step n+1), format control on input data provided by user, checking format of intermediate data. Missing but requested steps for ensuring the pipeline running (such as indexing reference) are added automatically if omitted.

Reproducibility and traceability

TOGGLe ensures that all experimental results are reproducible, documented as well as ready to be retrieved and shared. Indeed, results are organized in a structured tree of directories: all outputs are sorted into separate directories grouped by analyses type (sample or global analyses) and by workflow step (see supplementary figure 1).

All parameters, commands executed as well as software versions are kept in logs and a PDF report. Files such as the pipeline configuration and reference data are duplicated in the output folder, in which are also produced the specific scripts used for the analyses. The original input files, at the opposite to reference and configuration files, are not duplicated to reduce disk usage. Finally, a PDF report for the whole analysis is produced, providing global and visual informations for each sample at each step of the workflow. This report provides a diagram of the workflow, the parameters used (configuration file, softwares version,...) and summary statistics for the key files generated by the pipeline.

Error tracking and reentrancy

All errors and warnings encountered are recorded in logs; hence, finding the origin of an error is simplified. If a sample provokes an error, this sample will be ignored for the rest of the analysis, and the failing reported in the error logs. TOGGLe can also be relaunched after failing or when adding a new samples, and only failed steps or new samples will be re-computed.

Large numbers of sample analyzed

There is no true limits to the number of samples or the sequencing depth of a project that TOGGLe can take in charge. TOGGLe was already used on hundreds of samples jointly (up to 3,000 by now), from different types of assays (RNAseq, GBS, WGS,...), different analyses (polymorphism detection, read count), and on different organisms [9]. The only observed current limits are the data storage and the computing capacities available.

Documentation

Installation, quick and complete user manuals, screencasts and a complete developer documentation are available on our website http://toggle.southgreen.fr. In addition, we also provide pre-packed configuration files for different types of classical analyses.

Availability

The source code is freely available at http://toggle.southgreen.fr, under the double license GNU GPLv3/CeCiLL-C.

The TOGGLe website comprises a comprehensive step-by-step tutorial to guide the users to install and run the software. An online issue request is also available under GitHub of the South Green platform[8] (http://github.com/SouthGreenPlatform/TOGGLE/issues/) to report bugs and troubleshooting.

Bibliography 1 R. Blawid, J. Silva, and T. Nagata.
Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline.
Annals of Applied Biology, 170(3):301-314, 2017. 2 S. Djebali, V. Wucher, S. Foissac, C. Hitte, E. Corre, and T. Derrien.
Bioinformatics Pipeline for Transcriptome Sequencing Analysis, pages 201-219.
Springer New York, New York, NY, 2017. 3 J. Leipzig.
A review of bioinformatic pipeline frameworks.
Briefings in Bioinformatics, 18(3):530-536, 2017. 4 H. Li.
A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.
Bioinformatics (Oxford, England), 27(21):2987-93, nov 2011. 5 H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin.
The sequence alignment/map format and samtools.
Bioinformatics, 25(16):2078-2079, 2009. 6 A. I. Maarala, Z. Bzhalava, J. Dillner, K. Heljanko, and D. Bzhalava.
Virapipe: Scalable parallel pipeline for viral metagenome analysis from next generation sequencing reads.
Bioinformatics, page btx702, 2017. 7 P. Muir, S. Li, S. Lou, D. Wang, D. J. Spakowicz, L. Salichos, J. Zhang, G. M. Weinstock, F. Isaacs, J. Rozowsky, and M. Gerstein.
The real cost of sequencing: scaling computation to keep pace with data generation.
Genome Biology, 17(1):53, Mar 2016. 8 SouthGreen.
The south green portal: a comprehensive resource for tropical and mediterranean crop genomics.
Current Plant Biology, 7-8:6-9, 2016 Nov 2016. 9 C. Tranchant-Dubreuil, S. Ravel, C. Monat, G. Sarah, A. Diallo, L. Helou, A. Dereeper, N. Tando, J. Orjuela-Bouniol, and F. Sabot.
Toggle, a flexible framework for easily building complex workflows and performing robust large-scale ngs analyses.
bioRxiv, 2018. 10 X. Wu, T.-K. Kim, D. Baxter, K. Scherler, A. Gordon, O. Fong, A. Etheridge, D. J. Galas, and K. Wang.

srnanalyzer-a flexible and customizable small rna sequencing data analysis pipeline. *Nucleic Acids Research*, page gkx999, 2017.

Keywords: Data, intensive analysis, Workflow manager, parallel computing, reproducible research

Recipes for a successful collaborative software development: the test-case of bioconda. Application to the bioconvert: a bioinformatics format converter library.

Anne Biton [1], Bryan Brancotte [1], Yoann Dufresne [1], Thomas Cokelaer * [1], Kenzo-Hugo Hillion [1], Etienne Kornobis [1], Pierre Lechat [1], Rachel Legendre [1], Frédéric Lemoine [1,2], Blaise Li [1], Nicolas Maillet [1], Amandine Perrin [1], Bertrand Néron [1], Rachel Torchet [1], Nicolas Traut [3], Anna Zhukova [1,2]

[1] Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 25-28 Rue du Docteur Roux, 75015 Paris, France

[2] Unité Bioinformatique Evolutive, C3BI USR 3756, Institut Pasteur CNRS – Institut Pasteur de Paris – France

[3] Unité de Génétique Humaine et Fonctions Cognitives, Département de Neuroscience, Institut Pasteur, Paris, France – Institut Pasteur de Paris – France

"Data science is an interdisciplinary field of scientific methods, processes, algorithms, and systems to extract knowledge from data in various forms, either structured or unstructured, similar to data mining" (dixit wikipedia).
In bioinformatics, plethora of data formats exist, relating to different fields such as assembly, phylogeny, or systems biology to cite just a few. Although unstructured data are rare, the number of structured data is extremely large and diverse; they may be old, or have a complex syntax; they may be poorly documented or ambiguous; they may be standard or obsolete; they may be binary or human-readable. Moreover, as of today, there are more than 3000 bioinformatics software available on the bioconda website; many more are available on various websites such as github or other less visible websites. The interactions between so many tools require interoperability, which often involves data conversions. Scientists spend a significant amount of time in either understanding these formats or converting them to other formats.

Consequently, bioinformatic proficiency requires an increasing expertise in data format and data conversions, reducing the time and energy available for actual scientific investigation. Similarly, scientists from other fields need to devote lots of time to the understanding of these formats and the related conversions.

Here, we present a collaborative project available on github to help scientists to share recipes and conversions in a single solution, thus saving them from knowing several dozen software. Surprisingly, there are very few attempts at solving this problem in the bioinformatics community. Possibly because the task seems strenuous and challenging for a single individual. Indeed,

[*] Speaker

one needs to know many formats and software. The few attempts that have been made mainly resulted in tools that cover only a few format conversions, which lack testing or are not maintained anymore. This may seem surprising, given the potential usefulness of such framework in the community.

Many successful collaborative projects are available on github. One instance is the bioconda project (Nature methods, 3000 packages, 400 contributors). How such projects succeed in sharing and create synergy between scientists? We will try to answer to that question and tell the audience how such projects were created and evolved, and how they succeeded in communicating effectively.

In the second part of the talk, we will come back to the problem of data format conversion. More specifically, we will present the best practices of successful open source software that we followed (testing, one line documentation and openness), and how we applied the recipe used within the bioconda project to the bioconvert project. We will then show how with such techniques and little time and efforts, we can build an open source project from scratch. More importantly, we will show how we reduced the complexity of the format conversion problem using a plugin system, and how ready-to-go recipes will help new developers and users in integrating their own data conversion or methods.

We will also describe the current status of the bioconvert project: after only 6 months of existence, it includes 40 formats, 80 possible data conversions and has been downloaded 6300 times from pypi website. In fact many more conversions are available thanks to an implementation that allows conversions between formats for which no direct conversion plugin exist, via a succession of intermediate formats. With 15 collaborators and 800 commits, the project already provides a tool ready for production. Additionally, Bioconvert follows standards from software development with continuous integration on travis and automatic online documentation on readthedocs.

We will conclude this presentation with the future of the bioconvert project: full benchmarking to compare different conversions from various tools, ability to create web site, standalone application, biocontainers (e.g. singularity) and web services for programmatic online requests.

References:

> Documentation: http://bioconvert.readthedocs.io/en/master/
>
> Github : https://github.com/biokit/bioconvert

Bioconda: https://www.biorxiv.org/content/early/2017/10/27/207092 , http://blogs.nature.com/blog/tag/bio

Keywords: bioconda, software development, data conversion, interoperability, bioconvert

Genome-wide supervised learning of non-coding loci associated to complex phenotypes

Aitor Gonzalez *† [1], Marie Artufel [2,3], Pascal Rihet [4]

[1] Theories and Approaches of Genomic Complexity (TAGC) – Aix Marseille Université, Institut National de la Santé et de la Recherche Médicale : U1090, Centre National de la Recherche Scientifique – Théories et approches de la complexité génomiqueParc scientifique de Luminy163 avenue de Luminy13288 Marseille cedex 9, France

[2] Aix-Marseille Université, UMR1090 TAGC, Marseille, France – Aix-Marseille Université - AMU – France

[3] INSERM, UMR1090 TAGC, Marseille, France – Centre de Recherche Inserm – France

[4] Technologies avancées pour le génôme et la clinique (TAGC) – Inserm, Université de la Méditerranée - Aix-Marseille II – Parc scientifique de Luminy - 163 avenue de Luminy 13288 Marseille cedex 9, France

Genome-wide association studies (GWAS) link genetic loci to complex phenotypes in humans. Linkage disequilibrium (LD) blocks make it difficult to distinguish functional SNPs in associated loci. Therefore, functional SNPs are prioritized using causal SNP prediction tools and enriched with gene regulatory molecular markers such as DNA accesibility. However, priorization of GWAS SNPs is usually very low as measured by the AUC performance value.
We have trained a model called TAGOOS (TAG SNP bOOSting) to predict associated SNPs based on 4685 molecular variables such as chromatin marks and transcription factors and have computed scores genome-wide for intronic and intergenic regions. The TAGOOS scores enrich and prioritize unseen GWAS SNPs with higher performances than other available bioinformatics tools. To our knowledge, this is the first time where supervised learning is succesfully applied to non-coding loci associated to complex phenotypes. In this poster, we will present several insights gained from this model.

Keywords: genetics, single, nucleotide polymorphism, complex phenotypes, supervised classification, machine learning

*Speaker

†Corresponding author: aitor.gonzalez@univ-amu.fr

K-merator, an efficient design of highly specific k-mers for quantification of transcriptional signatures in large scale RNAseq cohorts.

Sébastien Riquier *† [1,2], Anne-Laure Bougé [1,2], Vy Nguyen *

[1,2], Benoit Guibert [1,2], Jérôme Audoux [3], Daniel Gautheret [4], Thérèse Commes‡ [1,2], Anthony Boureux§ [1,2]

[1] Cellules Souches, Plasticité Cellulaire, Médecine Régénératrice et Immunothérapies (IRMB) – Centre Hospitalier Régional Universitaire [Montpellier], CHU Saint-Eloi, Institut National de la Santé et de la Recherche Médicale : U1183, Université de Montpellier – Institute for Regenerative Medicine and Biotherapy - 80 rue Augustin Fliche 34295 Montpellier - Cedex 5, France

[2] Institut de Biologie Computationnelle (IBC) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut National de la Recherche Agronomique, Institut National de Recherche en Informatique et en Automatique, Université de Montpellier, Centre National de la Recherche Scientifique – 95 rue de la Galéra, 34095 Montpellier, France

[3] SeqOne – Institut National de la Santé et de la Recherche Médicale - INSERM – Hôpital Saint Eloi 80 rue Augustin Fliche 34295 MONTPELLIER - Cedex 5 FRANCE, France

[4] Institut de Biologie Intégrative de la Cellule (I2BC) – Université Paris-Sud - Paris 11, Commissariat à l'énergie atomique et aux énergies alternatives : DRF/I2BC, Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR9198 – Bâtiment 21, 1 avenue de la Terrasse, 91198 Gif/Yvette cedex, France

Following identification of candidate RNA biomarkers, their specificity has to be determined by

quantification in a large set of RNAseq data.

However, transcript quantification using classical approach, that relies on alignement methods, can not be easily adapted to the quantification of small set of candidate transcripts in large set of data.

The work of D.Gautheret and T.Commes groups on the DE-kupl pipeline (Audoux et al. 2017) has showed that k-mer decomposition constitute a new way to process RNA-Seq data for the identification of transcriptional signatures. Based on the same idea, k-mers can be used to quantify gene expression, transcript or any other transcriptional event level in a more specific, and less ressources-consuming way than classical approaches.

We present Kmerator, a tool for the design of specific tags based on the decomposition of transcript sequence into k-mers and the selection of a set of specific tags dedicated to transcript or gene expression mesurement.First we show that results of transcript expression

*Speaker

†Corresponding author: sebastien.riquier@inserm.fr

‡Corresponding author: Therese.Commes@inserm.fr

§Corresponding author: Anthony.Boureux@univ-montp2.fr

mesurement using kmerator generated tag count compared to the most commonly used RNA-seq quantification tool, Kallisto, are similar. Then we propose to use our strategy to set up a pipeline for RNA-seq data quality analysis. Using a particular set of tags we are able to predict metadata from public RNA-Seq data (Ribosomal depletion, orientation, Sex, Mycoplasma contamination).

Keywords: kmers, RNAseq

ELECTOR: EvaLuation of Error Correction Tools for lOng Reads

Camille Marchet *† [1], Pierre Morisse‡ [2], Lolita Lecompte [1], Antoine Limasset [3], Arnaud Lefebvre [2], Pierre Peterlongo [1], Thierry Lecroq [2]

[1] Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) – Universite de Rennes 1, Institut National de Recherche en Informatique et en Automatique – Avenue du général LeclercCampus de Beaulieu 35042 RENNES CEDEX, France

[2] Laboratoire dÍnformatique, de Traitement de lÍnformation et des Systèmes (LITIS) – Université de Rouen Normandie – Avenue de lÚniversité 76800 Saint-Étienne-du-Rouvray, France

[3] Université Libre de Bruxelles (ULB) – Avenue Franklin Roosevelt 50, 1050 Bruxelles, Belgium

Context

Pacific Biosciences (PB) and Oxford Nanopore (ONT) long reads, despite their high error rates and complex error profiles, were rapidly adopted for various applications. In particular, they are expected to help solving problems faced with short reads in the genomic assembly field. To overcome these high error rates, a plethora of error correction methods directly targeted at long reads were developed. These methods either aim at correcting the long reads solely based on the information contained on their sequences (self-correction), or use complementary short reads, relying on their important coverage depth and their low error rate (hybrid correction).

As the quality of the error correction has huge impacts on downstream processes, developing methods allowing to evaluate error correction tools with precise and reliable statistics is therefore a crucial need. However, works introducing new error correction methods usually evaluate the quality of their tools based on how the corrected long reads can be realigned to the reference. Despite being interesting, this information remains incomplete, and is likely not to mention poor quality reads, or regions to which it is difficult to align. In this work we propose ELECTOR, a novel tool that enables the evaluation of long read hybrid and self-correction methods, that provides relevant metrics and that scales to large datasets.

To date, LRCstats [1] was the only method able to realize long read correctors evaluation. LRCstats proposes a three-way alignment strategy that relies on pairwise alignments of both corrected and original versions of each read to the reference. LRCstats provides reads error rate before and after correction, as well as the detailed counts of every type of error. However, only studying the error rate of the reads is not a satisfying indication of the corrector's behaviour, as it does not report information about the putative insertions of new errors by the corrector. LRCstats is well-tailored for experiments with long reads of a few kilobases and for medium throughputs of less than 200 Mb. However we show that it can be more time and/or memory consuming than the actual error correction methods on larger experiments or reads longer than 10kb.

*Speaker

†Corresponding author: camille.marchet@irisa.fr

‡Corresponding author: pierre.morisse2@univ-rouen.fr

Contribution

In order to cope with these limits, we designed ELECTOR to 1/ compute more relevant metrics on long read correction; and 2/ scale to very long reads and large sequencing experiments. ELECTOR is directly compatible with a wide range of state-of-the-art error correction tools, without needing the user to perform any pre-processing. Therefore, it simply takes as input a set of reads, their corresponding corrected versions, and the corresponding reference genome. Contrary to LRCstats, it also includes additional steps performing and assessing corrected reads remapping and assembly, respectively using BWA mem [2] and Miniasm [3]. Output statistics include average identity of the alignments and genome coverage for the remapping part, and number of contigs, number of breakpoints, NGA50 and NGA75 for the assembly part. It is also meant to be a user-friendly tool, that delivers its results through different output formats, such as graphics than can be directly integrated to the users' projects.

First of all, the three-way alignment paradigm used in LRCstats is replaced by a multiple alignment of triplets of sequences in ELECTOR. Such an approach allows to efficiently compare the three different versions of each read: the uncorrected version, as provided by the sequencing experiment or by the reads simulator, the corrected version, as provided by the error correction method, and the reference version, that represents a perfect version of the original read, on which no error would have been introduced.

This choice of using a multiple alignment strategy allows ELECTOR to provide a wide range of metrics that assess the actual quality of the correction. In particular, ELECTOR is able to compute the recall, which is the rate of erroneous bases correctly modified (corrected) by the corrector, the precision, which measures the ability of the corrector not to add new erroneous bases, and the overall correct bases rate for each read. In addition to these classical metrics, ELECTOR also displays other results including GC content before and after correction, number of trimmed and/or split corrected reads, and mean missing size in those reads.

Secondly, we propose solutions to tackle scaling issues and thus to offer a faster and more scalable evaluation pipeline, which benefits large genomes processing. In particular, we coupled an implementation of multiple sequence alignment (MSA) using partial order graphs [4] to a seed strategy comparable to MUMmer [5] or Minimap [2]. This so-called *seed-MSA* strategy allows to divide the multiple sequence alignment problem, known to be time and memory consuming, into smaller instances. In addition to bringing an interesting methodological contribution, this allows to achieve a significant gain in resources footprint.

ELECTOR can be used on simulated data generated from state-of-the-art long reads simulation tools, such as NanoSim [6] or SimLoRD [7], on which introduced errors are precisely known, but also on real data. In the case of simulated data, the reference version of a given read is easily retrieved by parsing the files describing the introduced errors, generated by the simulator. In the case of real data, the reference sequences are retrieved by aligning the uncorrected reads to the reference genome, using Minimap2 [unpublished]. Only the best hit for each read is kept, and used to determine the corresponding reference sequence. In the case a read cannot align to the reference genome and thus cannot produce a reference sequence, this read is simply excluded from the analysis. In both cases, ELECTOR retrieves the reference versions of the reads by itself.

Results

Using bacterial and eukaryotic read sets, we validate our approach and demonstrate that 1/ our heuristic for multiple alignment of long reads provides results that are extremely similar to the original partial order graph alignment, while being several orders of magnitude faster, 2/ ELECTOR provides sound metrics in comparison to the state-of-the-art.

In order to validate our speedup strategy for multiple sequence alignment, we simulated two datasets from the *E. coli* genome, with SimLoRD. The first was composed of reads with a 1kb mean length, a 10% error rate and a coverage of 100X and the second was composed of reads with a 10kb mean length, a 15% error rate and a coverage of 100X. The reads from the two datasets were corrected with MECAT [8] with default parameters. The correction was then assessed both with MSA and *seed-MSA* strategies. Results of our experiments show that classic MSA and *seed-MSA* approaches only differ by a few digits in the presented metrics (recall, precision and correct bases rate of corrections). However, using *seed-MSA*, a substantial gain in time is achieved: while the classic MSA strategy has a subcubic runtime with respect to the read length and the average number of predecessors of nodes in the partial order graph, *seed-MSA* limits this drawback by working on small instances. As an example, for the second dataset, MSA and *seed-MSA* compute respectively a recall of 84.505% and 84.587%, a precision of 88.347% and 88.278%, and a correct bases rate of 95.290% and 95.250%, in 107 hours for the classical MSA approach, and in 42 minutes for the *seed-MSA* approach.

In order to validate the accuracy of ELECTOR's metrics, we then used SimLoRD to simulate three other datasets, respectively from *A. bayli, E. coli* and *S. cerevisiae.* Each of these datasets was composed of reads with a 8kb mean length, a 18% error rate, and a coverage of 20X. We corrected these datasets with various long read correctors, and assessed the correction results using both LRCstats and ELECTOR. Results of these experiments show that the metrics computed by ELECTOR are comparable to LRCstats outputs, but also allow us to highlight several novelties. For instance, on the long reads of the A. baylyi dataset corrected with Nanocorr [9], LRCstats reports an error rate of 0.005777 and ELECTOR reports a correct bases rate of 0.99534, which are in accordance, but ELECTOR only reports a recall of 0.97992, meaning that Nanocorr failed to correct 2% of the erroneous bases. Computation of these results is also more time-saving than LRCstats. In particular, on the *E. coli* dataset, LRCstats took an average of 3h50min to evaluate the quality of the correction of the different tools, while ELECTOR only took an average of 25 minutes.

Both hybrid and self correctors are included in this benchmark. Although our goal is not to debate on the comparison of correction methods efficiency, this is the first time several self-correctors appear assessed independently of a new error correction tool presentation.

Conclusion

We propose a novel and open-source method for fast long read correction assessment. Hybrid and self correctors are compatible. Our software ELECTOR outputs a wide range of metrics to finely understand the behavior of correction tools, even on large experiments. We compare ELECTOR to a previous work for correctors evaluation and show that it allows a faster and more extensive assessment of long read correction on several species.

Conclusions about pros and cons of hybrid vs self-correction and comparisons of correction paradigms vary a lot according to publications. At the moment, no coherent vision is proposed for long read correction and the field lacks a global study that goes over the sum of individual publications. ELECTOR could thus be a perfectly fitted basis for such benchmark study.

Availability

ELECTOR is an open-source software available on GitHub: github.com/kamimrcht/ELECTOR

References

Sean La, Ehsan Haghshenas, and Cedric Chauve. LRCstats, a tool for evaluating long reads correction methods. Bioinformatics, 33(22):3652-3654, 2017

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-

Wheeler transform. Bioinformatics, 26, 589–595.

Heng Li. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. Bioinformatics, 32(14):2103-2110, 2016

Christopher Lee, Catherine Grasso, and Mark F Sharlow. Multiple sequence alignment using partial order graphs. Bioinformatics, 18(3):452-464, 2002

Arthur L Delcher, Steven L Salzberg, and Adam M Phillippy. Using MUMmer to identify similar regions in large sequence sets. Current Protocols in Bioinformatics, pages 10-3, 2003

Chen Yang, Justin Chu, Rene L Warren, and Inanc Birol. NanoSim:

nanopore sequence read simulator based on statistical characterization. GigaScience, 6(4):1-6, 2017

Bianca K St́ocker, Johannes Ḱoster, and Sven Rahmann. SimLoRD: Simulation of long read data. Bioinformatics, 32(17):2704-2706, 2016

Chuan-Le Xiao, Ying Chen, Shang-Qian Xie, Kai-Ning Chen, Yan Wang, Yue Han, Feng Luo, and Zhi Xie. MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. Nature Methods, 14(11):1072, 2017

Goodwin, S., Gurtowski, J., Ethe-sayers, S., Deshpande, P., Schatz, M. C., and Mccombie, W. R. (2015). Oxford Nanopore sequencing, hybrid error correction and *de novo* assembly of a eukaryotic genome. Genome Research, 25, 1750–1756.

Keywords: third generation sequencing, pacific biosciences, nanopore, correction, evaluation

Stochastic optimization approach to large scale inferences of single molecule dynamics

François Laurent * [1,2], Jean-Baptiste Masson [3,4]

[1] Bioinformatics and Biostatistics Hub, C3BI, Pasteur Institute, CNRS USR 3756, 25-28 rue du Dr Roux, 75015 Paris, France – Institut Pasteur de Paris, Centre National de la Recherche Scientifique - CNRS – France
[2] Decision and Bayesian Computation, Pasteur Institute, CNRS UMR 3571, 25-28 rue du Dr Roux, 75015 Paris, France – Institut Pasteur de Paris, Centre National de la Recherche Scientifique - CNRS – France
[3] Décision et processus Bayesiens / Decision and Bayesian Computation – Institut Pasteur [Paris] – France
[4] Centre de Bioinformatique, Biostatistique et Biologie Intégrative – Institut Pasteur [Paris] – France

Time-resolved super-resolution microscopy techniques, e.g. sptPALM and uPAINT, let us observe individual biomolecules at the whole cell scale. These observed biomolecules are probes of their environment and recent progress in high density tagging enables us to map their mesoscopic dynamics and extract information on the biochemical interactions taking place inside living cells.
We introduce the TRamWAy open source project that brings together a range of analyses on single molecule data. Notably, TRamWAy samples single molecule data at various space and time scales and fits models of random walk dynamics, as a tool to capture the dynamical properties of the underlying biological processes. TRamWAy relies on Bayesian inference to address these complex information retrieval tasks.

Here, we investigate the molecular mechanisms that drive the assembly of enveloped RNA viruses at the plasma membrane of the host cell. In the context of the human immunodeficiency virus type 1 (HIV-1), we study the dynamics of the viral Gag protein, which produces virus-like particles (VLP).

To this aim, we infer time-varying spatial maps of physical parameters that capture the evolving dynamics of the Gag protein as VLPs are assembling. This implies maximizing a posterior function of thousands of variables from noisy and heterogeneous Gag displacements. We introduce a new **stochastic optimization** algorithm that combines **selective mini-batch sampling** with **dimensionality reduction** for efficient gradient computation, local **mean field approximation** of the likelihood and **regularizing priors**. This algorithm samples the posteriori probability distribution (of thousands of variables) in polynomial time.
We quantify the stochastic error and characterize the spatio-temporal coordinating effect of the viral RNA genome on the complete assembly process.

*Speaker

Keywords: stochastic optimization, mini, batch sampling, mean field approximation

A first step toward the personalized interactome: predicting the impact of mutations on proteins interaction profiles

Zacharie Ménétrier * , Quentin Ferré , Diogo Ribeiro [1], Andreas Zanzoni [2], Cécile Capponi [3], Lionel Spinelli [1], C. Brun† [4]

[1] Theories and Approaches of Genomic Complexity, U1090, Inserm- Aix-Marseille Université (TAGC) – Institut National de la Santé et de la Recherche Médicale - INSERM, Aix-Marseille Université - AMU – Parc Scientifique et Technologique de Luminy Case 928 13288 Marseille cedex 9, France
[2] Theories and Approaches of Genomic Complexity, U1090, Inserm- Aix-Marseille Université (TAGC) – Institut National de la Santé et de la Recherche Médicale - INSERM, Aix-Marseille Université - AMU – Parc Scientifique et Technologique de Luminy Case 928 13288 Marseille cedex 9, France
[3] Laboratoire dÍnformatique et Systèmes (LIS) – Aix Marseille Université : UMR7020, Université de Toulon : UMR7020, Centre National de la Recherche Scientifique : UMR7020 – Aix Marseille Université – Campus de Saint Jérôme – Bat. Polytech, 52 Av. Escadrille Normandie Niemen, 13397 Marseille Cedex 20, France
[4] Theories and Approaches of Genomic Complexity, U1090, Inserm- Aix-Marseille Université (TAGC) – Institut National de la Santé et de la Recherche Médicale - INSERM, Aix-Marseille Université - AMU – Parc Scientifique et Technologique de Luminy Case 928 13288 Marseille cedex 9 - France, France

Most biological functions are shaped by protein interactions. In this regard, the interactome is a powerful tool to understand the relationship between genotype and phenotype at the protein interaction level. Distinct proteinvariants lead to distinct phenotypic outcomes through "edgetic"perturbations in interactome networks, i.e. protein interaction modifications [**1**]. Indeed, protein variations or mutations can alter interactions in different ways [**1**]. Some of them may lead to the loss of all the protein's interactions, whereas others may lead to a gain or a loss of only some of their interactions. Such alterations will affect the interactome's edges (since they represent interactions) and are thus called "edgetic perturbations" [**1**].
As the topology of the interactome can be greatly modified by some edgetic perturbations and only subtlerly by some others, therefore possibly leading to different phenotypes, the *Predgetic*project aims to provide a computational workflow to study and predict the effect of protein mutations in an interactome, through the use of network analysis and machine learning.

In a first step of this project, inorder to transition from the average to the personalized interactome, we need to determine the edgetism level of all the mutated proteins of a specific individual. For this, we seek to create a specific score reflecting the global perturbation potential of each protein's interaction profile(= the set of all the interactions of a protein).Indeed, the reference interactome of an organism can then be tuned with the addition of a probability of interaction profile's perturbation on each node, based on the score calculated for each mutatedor variantprotein, therefore modifying its topology. Such modification of the interactome's topology between different conditions would be of great interest, allowing us to clarify the link

*Speaker

†Corresponding author: christine-g.brun@inserm.fr

between the interactome and the phenotype.

In order to provide a method that allows the prediction of the score of the impact of the mutations on a protein profile interaction, we are using machine learning approaches and a large dataset of curated interactions. We use the IntAct dataset which contains 22096 entries of protein mutations and their consequence (loss or conservation) on protein interactions **[10]**. However, this dataset is incomplete and does not give the full interaction profile for most mutations. Therefore, we first seek to complete the dataset through transductive methods, which consists in classifying specific cases (ie. determining the loss or conservation of interactions where the data is missing) based on the existing observed cases (known consequences of mutations), without the need to infer general rules **[11].**

For this endeavor, the protein sequences were first vectorized using the conjoint triads method [**4**]. Concatenated vectors were then classified using a Random Forest [**5**] algorithm, resulting in a 80% accuracy. We tested the limits of this method by restricting the provided information : using only information on the mutated protein, the classifier was still able to predict with an accuracy of around 80%; even though it should not have been possible to predict whether two proteins interact without information about both proteins involved. These attempts revealed a large homogeneity in the dataset, meaning most mutations either lead to total loss or total conservation of interactions.

Based on this newly highlighted need for specificity, methods relying on protein interfaces were tested. Protein interfaces were downloaded from *InteractomeInsider* [**6**], and a matrix of amino acid affinity was retrieved from the *Mechismo* [**7**] website. The difference in the distribution of the affinity score between wild-type and mutated proteins was not sufficient to differentiate between the aforementioned two classes of consequences of the mutations on the interactions (conservation or loss of the interaction).

To increase the visibility of the mutation signals, we sought to quantify their biological effects beyond the simple modification of sequences, by considering the mutation within its biological context; as such, we turned to word embedding, which is a feature learning technique where words are mapped to vectors of real numbers [**8**]. The embedding is realized in-context, and as such the vector contains information about the semantic meaning of the word instead of treating it as a raw sequence of letters. BioVec [**9**] provides such an embedding for amino acid trimers, trained through Swiss-Prot [**14**]. We use the BioVec representation to pre-encode our data so as to make the mutations more visible to our machine learning models. In order to emphasize the importance of the mutation and detect more abstract and composite representations, we turned to deep learning methods such as convolutional neural networks (CNNs) [**12**]and recurrent neural networks (RNNs) [**13**]. CNNs and RNNS were tested with trivial data and were able to perform pattern recognition in protein sequences. For now those models are still confounded by the proximity between wild-type and mutated proteins. We are currently amending the hyperparameters to allow the model to recognize the mutated patterns.

On a global perspective, we want to address the need for contextualized interactomes with the *Predgetic*workflow. The first step of this workflow is to provide a machine learning model that must predict the impact of mutations on a protein's global interaction profile. To use the IntAct dataset in this endeavor, transductive methods can be applied to complete the interaction profiles of mutations. However, a strong homogeneity in the IntAct dataset led us to revisit our approach. Even though, it would be interesting to discover if this homogeneity is the result of a curation bias or is the direct consequence of a biological fact. Future work will focus on the issue of predicting the mutations' interaction profiles through machine learning, which is still an open issue. Learning from pairs of features as in protein interactions also needs to be addressed in a generic method. Continuous improvement in external features acquisition is also pursued.

References

[**1**]*Zhong Q, Simonis N, Li Q-R, et al.* Edgetic perturbation models of human inherited disorders. Molecular Systems Biology. 2009;5:321. doi:10.1038/msb.2009.80.

[**2**] *Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature genetics. 1999; 22(3): 231–8.

[**3**] *Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001; 409(6822): 928–33.

[**4**]*Shen J, Zhang J, Luo X, et al.* Predicting protein–protein interactions based only on sequences information. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(11):4337-4341. doi:10.1073/pnas.0607879104.

[**5**]*Tin Kam Ho.* 1995. Random decision forests. In Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1 (ICDAR '95), Vol. 1. IEEE Computer Society, Washington, DC, USA, 278-.

[**6**] Interactome INSIDER: a structural interactome browser for genomic studies. *MJ Meyer, JF Beltrán, S Liang, R Fragoza, A Rumack, J Liang, X Wei, H Yu*- Nature Methods, 2018

[**7**]*Betts MJ, Lu Q, Jiang Y, et al.*Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. Nucleic Acids Research. 2015;43(2):e10. doi:10.1093/nar/gku1094.

[**8**]Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean

arXiv:1310.4546

[**9**] *Asgari E, Mofrad MRK.* Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. Kobeissy FH, ed. PLoS ONE. 2015;10(11):e0141287. doi:10.1371/journal.pone.0141287.

[**10**]*PORRAS MILLÁN, Pablo et al.*The MINTAct Archive for Mutations Influencing Molecular Interactions. Genomics and Computational Biology, [S.l.], v. 4, n. 1, p. e100053, dec. 2017. ISSN 2365-7154. Available at: . Date accessed: 07 may 2018. doi: https://doi.org/10.18547/gcb.2018.vol4.iss1.e

[**11**]Learning by Transduction

Alex Gammerman, Volodya Vovk, Vladimir Vapnik

arXiv:1301.7375

[**12**] *Toshiteru Homma*, Les Atlas et Robert II Marks, " An Artificial Neural Network for Spatio-Temporal Bipolar Patters: Application to Phoneme Classification ", Advances in Neural Infor-

mation Processing Systems, vol. 1, 1988, p. 31–40

[13]*Schmidhuber, J´urgen*(1993). Habilitation thesis: System modeling and optimization

[14] *Bairoch A, Apweiler R.*The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research. 2000;28(1):45-48.

you can use ColWiz google docs plugin for the references

ok for now i will be using brute force method but next time if you show me how to use it i will test the plugin

Keywords: Machine learning, protein interaction, interactome, edge tics

ThaliaDB, a tool for data management and genetic diversity data exploration

Delphine Steinbach *† 1

1 INRA (GQE-Le Moulon) – Institut National de la Recherche Agronomique – Ferme du moulon, 91190 Gif sur Yvette, France

Diversity and association genetics studies lead to manipulate a large number of individual, lines, clones and/or populations. Moreover, emergence of high-throughput technologies for both genotyping and phenotyping generates a large amount of data. These data need to be stored and managed in order to make requests and to organize datasets to be able to perform genetic diversity data exploration and association genetics analysis. The new version of ThaliaDB, V3.2, is developed for scientists to facilitate their data management and analysis. The database holds genetic resources data (germplasm/accessions), seed lots, samples, markers and genotyping and phenotyping datasets (fields environments, multiple traits and conditions). It is well adapted for data, useful to apply GWAS or genomic selection methods. It can manage high-throughput results coming from different projects and experiments and propose several views and options to explore these data and to give access to them for reuse. This Web tool offers to users a Select (Data view) mode and an Admin (Data administration and loading) mode. Data confidentiality is maintained using user accounts and specific levels of rights can be set on data. It enables data extraction in CSV format. A version exists today in our lab with maize data produced from projects of A. Charcosset's GQMS team and theirs partners. Perspectives are to use it for tomato, wheat and poplar data. The software is currently in improvement with funding of Amaizing, Investment for the future, project. It is developed in Python under Framework Django, running under PostGreSQL and MongoDb databases management systems. Contact: delphine.steinbach@inra.fr for more information and collaboration.

Keywords: Database, plant, GWAS, genotyping, phenotyping, highthrouput, diversity analysis

*Speaker

†Corresponding author: delphine.steinbach@inra.fr

Exploration of EDAM ontology and bioinformatics resources in a reusable web-visualization

Bryan Brancotte *† [1], Christophe Blanchet [2,3], Hervé Ménager‡ [1]

[1] Bioinformatics and Biostatistics Hub of the C3BI, Institut Pasteur – Institut Pasteur de Paris – 25 rue du Dr Roux. 75015 Paris, France
[2] Institut de Biologie et Chimie des protéines (IBCP) – CNRS : FR3302, Université de Lyon – France
[3] IFB-CORE – Inserm : US21, CNRS : UMS3601, Université Paris XI - Paris Sud – 1 avenue de la Terrasse - Bâtiment 21 - 91190 Gif/Yvette, France

Labelling, indexing and describing a Bioinformatics resource, whether it is a software, a database, or a service is of a great help when it comes to promoting it to various user communities. As an example, the ELIXIR bio.tools [3] registry contains more than ten thousands software and service entries. In this context, the use of controlled vocabularies to describe the resources is of a paramount importance. In bio.tools, this need is addressed by the EDAM Ontology [2], which proposes a controlled vocabulary hierarchically organized around four axes which describe types of data, formats, operations and topics.
We here present the EDAM Browser, a client-side web-based visualization javascript widget that provides an interface to navigate EDAM. This browser is tailored to the needs of EDAM users who might not be ontology experts. It can, among other things, be used to help describing resources, and to facilitate and foster community contributions to EDAM. The EDAM Browser allows users to explore it with an interface tailored to its structure and properties. Its interface is not designed to be a generic ontology navigation and edition platform, a goal already achieved by many other systems such as AberOWL[1], BioPortal[6], OLS - Ontology Lookup Service[4], Ontobee[7] and WebProtégé[5].

Rather, it aims at providing features requested by most users and contributors, which we detail below.

## Availability and re-usablility

The EDAM browser is available publicly and anonymously at https://ifb-elixirfr.github.io/edam-browser/. In addition to this, its lightweight architecture makes it easy to download and run on any server or personal computer, either as a local HTML file or on a web server. It is possible to integrate the EDAM Browser and its tree representation in external websites and applications, providing a simple way for third party websites to promote EDAM-labeled resources. Both the autocomplete input field and the tree visualization are re-usable: a demonstration code is available at https://ifb-elixirfr.github.io/edam-browser/demo.html , showing how the tree can

*Speaker
†Corresponding author: bryan.brancotte@pasteur.fr
‡Corresponding author: hmenager@pasteur.fr

be integrated, how the user can interact with the tree, and how to programmaticaly interact with the tree in JavaScript.

## Information display

As much as possible, the user interface aims at simplicity and relevance to the specific domain of EDAM. The creation of an interface that displays all of the information necessary to users, and avoids the use of ontology development jargon is a major goal of this project.

We also take into account the specificities of the structure of EDAM: while being represented as a tree, it is in fact a directed acyclic graph, meaning that a term can have more than one parent. In order to improve readability when a term is selected (1) all the term's positions are shown; and (2) all paths from the root node are highlighted. A good example of this display is the Phylogeny topic which can be seen here https://ifb-elixirfr.github.io/edam-browser/#topic_0084 .

The interface also permits the navigation between different axes of the ontology, based on the EDAM properties that define their relationships (e.g. this "format" represents this type of "data", this "data" is an output of this "operation" or is specific of this "topic"). One last salient feature of the interface is the representation of the usage of the selected concept in annotated resource collections, such as bio.tools, BioSphere, BioWeb and TeSS.

## Performance and flexibility

One of the specificities of EDAM is its relatively small size in comparison with large ontologies like Gene Ontology. This reduced size makes it easy to load entirely the contents to be displayed in the browser's memory, and enables a very fast navigation, with no need to rely on server calls during this navigation (except for displaying usage statistics from external annotated resources).

Using the EDAM Browser to explore a local or in-development version is possible. The loaded file should be formatted as a JSON file following the schema accessible at https://ifb-elixirfr.github.io/edam-browser/ontology.schema.json. The edam2json utility can be used to generate the ontology in this format from any EDAM owl file (available at https://github.com/edamontology/edam2json).

An ontology is loaded into the EDAM Browser by clicking on the button labelled "Custom" at the top of its interface, and specifying either a public URL to the file or a local path to load it from.

## Ease of community contributions.

Letting users easily propose changes to the ontology improves its acceptance by the community, as well as its long term maintainability. To facilitate these suggestions, the EDAM Browser lets users access a form letting them propose changes at any point of their exploration. These suggestions are automatically formatted as github issues ready to be submitted by the user.

## References

R. Hoehndorf, L. Slater, P. N. Schofield, and G. V. Gkoutos. Aber-owl: a framework for ontology-based data access in biology. BMC bioinformatics, 16(1):26, 2015.

J. Ison, M. Kalaš, I. Jonassen, D. Bolser, M. Uludag, H. McWilliam, J. Malone, R. Lopez, S.

Pettifer, and P. Rice. Edam: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics, 29(10):1325–1332, 2013.

J. Ison, K. Rapacki, H. Ménager, M. Kalaš, E. Rydza, P. Chmura, C. Anthon, N. Beard, K. Berka, D. Bolser, et al. Tools and data services registry: a community effort to document bioinformatics resources. Nucleic acids research, 44(D1):D38–D47, 2015.

S. Jupp, T. Burdett, C. Leroy, and H. E. Parkinson. A new ontology lookup service at embl-ebi. In SWAT4LS, pages 118–119, 2015.

T. Tudorache, C. Nyulas, N. F. Noy, and M. A. Musen. Webprotégé: A collaborative ontology editor and knowledge acquisition tool for the web. Semantic web, 4(1):89–99, 2013.

P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic acids research, 39(suppl 2):W541–W545, 2011.

Z. Xiang, C. Mungall, A. Ruttenberg, and Y. He. Ontobee: A linked data server and browser for ontology terms. In ICBO, 2011.

Keywords: javascript, bioinformatics, d3js, d3.js, ontology, visualization, edamontology, ontology

IMGT® genomic annotation of the dog (Canis lupus familiaris) seven immunoglobulin (IG) or antibody and T cell receptor (TR) loci

Imène Chentli *† [1], Perrine Pégorier * ‡ [1], Saïda Saljoqi§ [1], Géraldine Folch [1], Joumana Michaloud [1], Véronique Giudicelli [1], Patrice Duroux [1], Sofia Kossida¶ [1], Marie-Paule Lefranc‖ [1]

[1] IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM) – Institut de Génétique Humaine (IGH) CNRS Université de Montpellier UMR 9002 – 141 rue de la Cardonille 34396 Montpellier Cedex 5, France

INTRODUCTION

IMGT®, the international ImMunoGeneTics information system®, http://www.imgt.org [1], is the global reference in immunogenetics and immunoinformatics [2], founded in 1989 by Marie-Paule Lefranc at Montpellier (Université de Montpellier and CNRS). IMGT® is a high-quality integrated knowledge resource specialized in the immunoglobulins (IG) or antibodies, T cell receptors (TR), major histocompatibility (MH) of human and other vertebrate species, and in the immunoglobulin superfamily (IgSF), MH superfamily (MhSF) and related proteins of the immune system (RPI) of vertebrates and invertebrates.

The genome of the vertebrates with jaws (*Gnathostomata*), which appeared in the evolution about 450 million years ago, includes the IG, TR and MH genes characteristic of the adaptive immune repertoires. Currently, there are 244 annotated vertebrate genomes including 112 from mammals at NCBI.

In humans and other mammals, there are seven main *loci* for IG and TR: three for IG (IGH, IGK and IGL) and four for TR (TRA, TRB, TRD and TRG). IMGT® genomic annotated data are classically displayed in IMGT Repertoire Web Resources (Locus description, Locus representation, Gene tables, Alignments of alleles). So far the number of species present in the IMGT Web Resources reaches 40, however only two species, *Homo sapiens* and *Mus musculus,* have been fully annotated for their seven antigen receptor *loci.* This biocuration has been perfomed manually and the standardized annotation has allowed data entry in IMGT® databases

[*] Speaker

[†] Corresponding author: imene.chentli@igh.cnrs.fr

[‡] Corresponding author: perrine.pegorier@igh.cnrs.fr

[§] Corresponding author: saida.saljoqi@igh.cnrs.fr

[¶] Corresponding author: sofia.kossida@igh.cnrs.fr

[‖] Corresponding author: marie-paule.lefranc@igh.cnrs.fr and tools [2]. The dog (*Canis lupus familiaris*) represents the first species for which the seven *loci* are annotated simultaneously. The biocuration was performed on the *loci* extracted from genome assembly.

METHODOLOGY

The seven IG and TR *loci* of the dog genome were recently described [3]. *CanFam3.1* is the last assembly (March 2015) of a female boxer (GenBank assembly accession: GCA_000002285.2, GenBank BioProject accession: PRJNA13179). Each *locus* sequence was localized on the corresponding chromosome and extracted. As the *locus* orientation on a chromosome can be either forward (FWD) or reverse (REV), the REV *locus* sequences were placed in the 5' to 3' *locus* orientation. Each *locus* sequence was assigned to a unique IMGT® accession number (IGH: IMGT000001, IGK: IMGT000002, IGL: IMGT000003, TRA/TRD: IMGT000004, TRB: IMGT000005 and TRG: IMGT000006).

RESULTS

The dog **IGH** *locus*, on chromosome 8 (REV), spans 1425 kilobases (kb) and consists of 89 IGHV genes belonging to 4 IGHV subgroups (36 functional, 2 ORF and 51 pseudogenes), 6 IGHD genes (5 functional and 1 ORF), 6 IGHJ genes (5 functional and 1 ORF) and 5 IGHC genes (4 functional and 1 ORF). The dog **IGK** *locus*, on chromosome 17 (REV), spans 349 kilobases (kb) and consists of 22 IGKV genes belonging to 5 IGKV subgroups (13 functional, 1 ORF and 8 pseudogenes), 5 IGKJ genes (4 functional and 1 ORF) and 1 IGKC gene (functional). The dog **IGL** *locus*, on chromosome 26 (FWD), spans 2583 kilobases (kb) and is still in progress. In June 2018, 261 IGLV genes belonging to 7 IGLV subgroups (69 functional, 12 ORF and 180 pseudogenes), 9 IGLJ genes (functional) and 9 IGLC genes (functional) were identified.

The dog **TRA** *locus*, on chromosome 8 (FWD), spans 743 kilobases (kb) and consists of 56 TRAV genes (34 functional and 22 pseudogenes) belonging to 30 TRAV subgroups, 59 TRAJ genes (40 functional, 12 ORF and 7 pseudogenes) and 1 TRAC gene (functional). The dog **TRB** *locus*, on chromosome 16 (REV), spans 271 kilobases (kb) and consists of 36 TRBV genes (22 functional, 1 ORF and 13 pseudogenes) belonging to 25 TRBV subgroups, 2 TRBD genes (functional), 12 TRBJ genes (9 functional, 2 ORF and 1 pseudogene) and 2 TRBC genes (functional). Like in the human (*Homo sapiens*) TRB *locus*, the functional TRBV30 gene is in inverted orientation of transcription downstream of the TRBC2 gene and rearranges by a mechanism of inversion. The dog **TRD** *locus*, on chromosome 8 (FWD), spans 344 kilobases (kb) and is embedded in the TRA *locus* between the TRAV and the TRAJ genes. It consists of 5 TRDV genes (3 functional belonging to 3 subgroups; 1 ORF and 1 pseudogene interspersed among the TRAV genes), 2 TRDD genes (functional), 4 TRDJ genes (3 functional and 1 ORF), and 1 TRDC gene (functional). Like in the human TRD *locus*, the functional TRVD3 gene is in inverted orientation of transcription downstream of the TRDC gene and rearranges by a mechanism of inversion. The dog **TRG** *locus*, on chromosome 18 (FWD), spans 447 kilobases (kb) and consists of 16 TRGV genes (8 functional and 8 pseudogenes) belonging to 7 TRGV subgroups, 16 TRGJ genes (7 functional, 3 ORF and 6 pseudogenes) and 8 TRGC genes (6 functional, 1 ORF and 1 pseudogene).

CONCLUSION AND PERSPECTIVES

Information on the IMGT® gene and *locus* reference sequences is available in the classical IMGT Repertoire pages, to which were added two novel pages: Locus in genome assembly and Locus gene order (with links to Locus representation). The annotation of the seven dog *loci* gives access to the study and comparison of the expressed adaptive immune repertoires in veterinary, normal and pathological situations, using IMGT® tools such as, for nucleotide sequence analysis, IMGT/V-QUEST [4], for high throughput next generation sequencing (NGS), IMGT/HighV-QUEST [5], and for domain amino acid sequence analysis, IMGT/DomainGapAlign [6]. The curated IG and TR dog genes and alleles have been entered in the IMGT/GENE-DB database [7] and the corresponding IMGT® reference directories [1] will be used for coherent gene and sequence annotations of IG and TR *loci* in other *Canidae*. Indeed, this simultaneous biocuration of the seven IG and TR *loci* in *Canis lupus familiaris* has extended, between *loci*, the IMGT® concepts of classification (nomenclature) and description (labels) for a standardized IMGT® genomic annotation at the chromosome and genome assembly levels.

Dogs are an excellent model for human disease. For example, the treatment of canine lymphoma has been predictive of the human response to that treatment [3]. Study of the antigen receptor and immune response in dogs offers a unique opportunity for potential applications in veterinary and human medicine.

REFERENCES

Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, Hadi-Saljoqi S, Sasorith S, Lefranc G, Kossida S. IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucl. Acids Res. 2015 Jan;43(Database issue):D413-22. doi: 10.1093/nar/gku1056. Epub 2014 Nov 5 Free Article. PMID: 25378316

Lefranc M-P. Immunoglobulin (IG) and T cell receptor genes (TR): IMGT® and the birth and rise of immunoinformatics. Front Immunol. 2014 Feb 05;5:22. doi: 10.3389/fimmu.2014.00022. Open access. PMID: 24600447

Martin J, Ponstingl H, Lefranc M-P, Archer J, Sargan D, Bradley A. Comprehensive annotation and evolutionary insights into the canine (*Canis lupus familiaris*) antigen receptor loci. Immunogenetics. 2018 Apr; 70(4):223-236. doi: 10.1007/s00251-017-1028-0. Epub 2017 Sep 19. Free article. PMID: 28924718

Brochet X, Lefranc M-P, Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucl. Acids Res, 36, W503-508 (2008); doi:10.1093/nar/gkn316. PMID: 18503082

Alamyar E, Giudicelli V, Shuo L, Duroux P, Lefranc M-P. IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Res. 2012, April 20;8:1:2. doi: 10.4172/1745-7580.1000056. PMID: 22647994

Ehrenmann F, Kaas Q, Lefranc M-P. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. Nucl. Acids Res., 38, D301-307 (2010). Epub 2009 Nov 9; doi:10.1093/nar/gkp946. PMID: 19900967

Giudicelli V, Chaume D, Lefranc M-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucl. Acids Res., 33, D256-D261 (2005). PMID: 15608191

Keywords: IMGT, immunoinformatics, immunogenetics, immunoglobulin, T cell receptor, Canis lupus familiaris

Capra hircus and Ovis aries IGK loci simultaneous annotation in IMGT®

Viviane Nguefack Ngoune *† 1, Morgane Bertignac * ‡ 1, Géraldine Folch 1, Joumana Michaloud 1, Sofia Kossida§ 1, Marie-Paule Lefranc¶ 1

1 IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM) – Institut de Génétique Humaine (IGH) CNRS Université de Montpellier UMR 9002 – 141 rue de la Cardonille 34396 Montpellier Cedex 5, France

INTRODUCTION

IMGT®, the international ImMunoGeneTics information system®, http://www.imgt.org [1], is the global reference in immunogenetics and immunoinformatics [2], founded in 1989 by Marie-Paule Lefranc at Montpellier (Université de Montpellier and CNRS). IMGT® is a high-quality integrated knowledge resource specialized in the immunoglobulins (IG) or antibodies, T cell receptors (TR), major histocompatibility (MH) of human and other vertebrate species, and in the immunoglobulin superfamily (IgSF), MH superfamily (MhSF) and related proteins of the immune system (RPI) of vertebrates and invertebrates.

The genome of the vertebrates with jaws (*Gnathostomata*), which appeared in the evolution about 450 million years ago, includes the IG, TR and MH genes characteristic of the adaptive immune repertoires [2]. In humans and other mammals, there are seven main loci for IG and TR: three for IG (IGH, IGK and IGL) and four for TR (TRA, TRB, TRD and TRG). IMGT® genomic annotated data are classically displayed in IMGT Repertoire Web Resources (Locus description, Locus representation, Gene tables, Alignments of alleles).

The IG are B cell antigen receptors, expressed at the membrane of the B cells or secreted by plasma cells, and characterized by the huge diversity of their binding specificities. Classically IG proteins comprise two identical heavy chains (H) associated with two identical light chains (L) which belong, in higher vertebrates, to two chain types, kappa (IGK) or lambda (IGL). Several variable (V), diversity (D) (only present in the IGH locus), joining (J) and constant (C) genes compose the IGH, IGK, and IGL loci. The IG biosynthesis requires the recombination of these genes [2] and it is the result of these complex mechanisms of V-(D)-J rearrangements and junctional N-diversity and, for the IG, somatic mutations which creates the IG high diversity.

The IG loci have so far been explored in only a limited number of species. Indeed these loci are difficult to annotate owing to their multigene organization of highly similar genes and their biocuration requires a reliable and high quality locus assembly. The genomes of different ruminant species start becoming available, and among them, those of the domestic goat (*Capra*

*Speaker

†Corresponding author: viviane.nguefack-ngoune@igh.cnrs.fr

‡Corresponding author: morgane.bertignac@igh.cnrs.fr

§Corresponding author: sofia.kossida@igh.cnrs.fr

¶Corresponding author: marie-paule.lefranc@igh.cnrs.fr *hircus*) and of the sheep (*Ovis aries*).

METHODOLOGY

The San Clemente Island goat (*Capra hircus*) genome has been sequenced (GenBank BioProject accession: PRJNA290100) and the information related to the identification of the goat IGK locus is available on public databases [3]. The goat IGK locus is situated on chromosome 11 in forward (FWD) orientation (GenBank: NC_030818.1; genome assembly: ARS1). An IMGT flat file was created, IMGT000009, which comprises the extracted region 46500000-46946647 bp from the chromosome 11.

In order to localize the IGK locus in sheep (*Ovis aries*), the sequence of the most 5' V gene and of the most 3' C gene from the goat locus were used to BLAST the sheep genome (genome assembly: Oar_v4.0). The sheep IGK locus is situated on chromosome 3 in reverse (REV) orientation. An IMGT flat file was created, IMGT000010, which comprises the extracted region (complement) 59150070-59300000 bp from chromosome 3 in order to place the sequence in the 5' to 3' orientation of the locus.

The annotation tool IMGT/LIGMotif [4] was used for the identification of the variable (V), joining (J) and constant (C) genes. The biocuration and description of the identified genes were performed using IMGT® tools (IMGT/Automat, IMGT/NTItoVALD). All the labels (V-REGION, J-REGION, C-REGION...) of each type of genes (V, J and C genes) were characterized. Subsequently, a multiple alignment and a phylogenetic tree of the V-REGION sequences of the goat and sheep IGKV genes and of a representative of each IGKV subgroup in human were performed.

RESULTS

The goat (*Capra hircus*) IGK locus on chromosome 11 (FWD) spans 447 kilobases and consists of 21 IGKV genes (5 genes are functional, 3 genes are ORF and 13 genes are pseudogenes), 4 IGKJ genes (1 gene is functional and 3 genes are ORF) and 1 IGKC gene (functional).

The sheep (*Ovis aries*) IGK locus on chromosome 3 (REV) spans 150 kilobases and consists of 18 IGKV genes (5 genes are functional, 1 gene is ORF and 12 genes are pseudogenes), 4 IGKJ genes (1 gene is functional and 3 genes are ORF) and 1 IGKC gene (functional).

As expected, in more than 85% (18/21) of the cases, the V-REGION of goat and sheep are similar to each other. The same subgroups are found in both species: IGKV1, IGKV2, IGKV3, IGKV6 and a new subgroup was introduced for both species: IGKV8, which until now did not exist in the other species studied within IMGT®. Interestingly, beyond this similarity of the subgroups, the IGKV gene positions of both species are the same. A unique and common nomenclature of those genes was proposed.

Some degenerated IGKV genes which could not be assigned to a given subgroup were assigned to the same clan, IGKV(II), because these genes were closely related (on a phylogenetic tree) to *Homo sapiens* IGKV2, IGKV3, IGKV4 and IGKV6 subgroup genes, http://www.imgt.org/IMGTindex/Clan.pl . They include IGKV(II)-2, IGKV(II)-13, IGKV(II)-16 and IGKV(II)-19 for both species and IGKV(II)-10, IGKV(II)-12 only for goat.

The CDR-IMGT lengths [2], which structurally define the IG and TR V-REGION of the germline genes (http://www.imgt.org/IMGTScientificChart/Nomenclature/IMGT-FRCDRdefinition.html) and are visible in the IMGT Colliers de Perles structure [5], were used to characterize all the functional V-GENE of each IGKV subgroup in both species, designated in this study as Caphir and Oviari (using the 6-letter code IMGT taxon abbreviation). Three IGKV1 genes, Caphir IGKV1-1, IGKV1-6, and IGKV1-7 and Oviari IGKV1-1, IGKV1-6, and IGKV1-7, have the same [6.3.7] CDR-IMGT lengths. The IGKV2-8, IGKV2-9, IGKV2-14 and IGKV2-15 genes in both species have also the same [11.3.7] CDR-IMGT lengths, even if the functionality differs from goat to sheep for the IGKV2-9 and IGKV2-15 (pseudogenes in the sheep).

CONCLUSIONS AND PERSPECTIVES

The IMGT-ONTOLOGY axiom and concepts of CLASSIFICATION provide the rules for the IG gene classification [2]. Although *Homo sapiens* genes and subgroups are commonly used in IMGT® for the biocuration of mammalian species, it is the first time that within IMGT® the same type of locus for two relatively close species of ruminants was studied simultaneously. Owing to the multigene family structure of the IG locus, and the post-speciation locus evolution, no attempt was done so far for identification of orthologous V genes between species. However, the goat and sheep species are sufficiently close to each other in evolution that, for the first time, a unique nomenclature for orthologous V genes could be proposed for the two species. This study was possible owing to the V-REGION high similarity of IGKV genes which, at the same time, occupy similar gene order in the IGK locus. An extension of that approach is currently evaluated and underway for the IGL locus of goat and sheep within IMGT®. Standardization of the gene nomenclature of orthologous genes of IG and TR multigene families between closely related species will accelerate the exploration of the antigen receptor loci from novel genomes in veterinary ruminant species.

Beyond research, goat and sheep small ruminant species have a highly economic interest because they contribute significantly to the nutrition and cash income for many farmer in developing world such as Africa and South Asia. unfortunately both species are regularly affected by different strains of Morbillivirus, which are extremely contagious and lethal. The characterization of the adaptative immune responses in these species would be beneficial for vaccine and immune reagent development. [6]

REFERENCES

Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, Hadi-Saljoqi S, Sasorith S, Lefranc G, Kossida S. IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucl. Acids Res. 2015 Jan;43(Database issue):D413-22. doi: 10.1093/nar/gku1056. Epub 2014 Nov 5 Free Article. PMID: 25378316

Lefranc M-P. Immunoglobulin (IG) and T cell receptor genes (TR): IMGT® and the birth and rise of immunoinformatics. Front Immunol. 2014 Feb 05;5:22. doi: 10.3389/fimmu.2014.00022. Open access. PMID: 24600447

Schwartz JC, Philp RL, Bickhart DM, Smith TPL, Hammond JA. The antibody loci of the domestic goat (*Capra hircus*). Immunogenetics 2017 Oct 23., doi:10.1007/s00251-017-1033-3. PMID: 29063126

Lane, J., Duroux, P. and Lefranc, M.-P. From IMGT-ONTOLOGY to IMGT/LIGMotif: the IMGT® standardized approach for immunoglobulin and T cell receptor gene identification and description in large genomic sequences. BMC Bioinformatics 2010, 11:223; doi:10.1186/1471-2105-11-223. PMID: 20433708

Kaas, Q., Ehrenmann, F. and Lefranc, M.-P. IG, TR, MHC, IgSf and MhcSF: what do we learn from the IMGT Colliers de Perles? Brief. Funct. Genomic Proteomic, 2007, 6, 253-264. Epub 2008 Jan 21. doi:10.1093/bfgp/elm032. PMID: 18208865

Naveen Kumar, Sunil Maherchandani, Sudhir Kumar Kashyap, Shoor Vir Singh, Shalini Sharma, Kundan Kumar Chaubey, Hinh Ly. Peste Des Petits Ruminants Virus Infection of Small Ruminants: A Comprehensive Review. Viruses. 2014 Jun; 6(6): 2287–2327. doi: 10.3390/v6062287. PMID: 24915458

Keywords: IMGT, immunoinformatics, immunogenetics, immunoglobulin, T cell receptor, Capra hircus, Ovis aries

CeSGO : integrating data and project lifecycle on a bioinformatics facility

Cyril Monjeaud * [1], Matéo Boudet [1], Olivier Collin [1]

[1] GenOuest – Universté de Rennes, L'Institut National de Recherche en Informatique et e n Automatique (INRIA), Centre National de la Recherche Scientifique - CNRS, IRISA/OBELI – GenOuest core facility Univ Rennes, Inria, CNRS, IRISA F-35000 Rennes, France, France

Bioinformatics core facilities have to develop new services to support the growing needs of the Life Sciences community. Adopting a vision centered on the scientific project lifecycle and the scientific data lifecycle, the CeSGO (Centre e-Science Grand Ouest) project is developing science gateway services on top of a bioinformatics infrastructure. The combination of a collaborative environment with a computing infrastructure allows users to manage both scientific projects and scientific data in a simple and integrated way.
The main idea of CeSGO is to integrate as much as possible of the scientist task flow. To reach this goal, the different steps of the scientific project life cycle and the scientific data lifecycle were considered. For each step, a service was implemented according to its characteristics and specificities in terms of data and user interface requirements. For tasks related to the project life cycle, there is mostly a need for collaborative environments to interact, discuss, write and share documents, plan. For tasks related to the scientific data lifecycle, there is a need for tools that can help scientists to manage, analyze and share their data.

The federation of collaborative tools and services in conjunction with a computing environment offers creates a gateway suitable for mid-size computing facilities. The fact that the data are co-localized with the collaborative services creates a "one-stop-shop" that will help scientists to adopt new usages. This on-premises environment makes it easier to fight against the silo effect induced by the use of too many unrelated services.

CeSGO collaborative environment

Collaboration

Collaboration is a social network platform. Users can create, join scientific groups and exchange inside them. All members of a group can edit documents and upload data files. A forum and event management are available in each group.

This platfom allow users to create and manage a custom WordPress website for their projects

Projects

This service is based on Kanban, a visual system for managing tasks within a project. It

*Speaker

manages private or shared projects with a group of users. Calendar and Gantt diagram are available to handle tasks in addition to the board view.

Instant

This web solution allows user to chat directly with another user or with a group of users. This service is complementary to CeSGO Collaboration portal.

CeSGO data management environment

Data-access

This service provides files access through a web interface. All files can be shared as links and made available to other users and groups. Synchronization is possible using a desktop client. The addition of OnlyOffice offers online and concurrent editing functionalities.

An in-house Application Programming Interface (API) was developed to improve interactions between Data-access and the GenOuest platform user's directories.

Research sharing

This service is a web-based resource for sharing scientific research datasets, is used. The datasets can be organized into investigations, studies and assays (biological experiments) based on the ISA (Infrastructure, Study, Assay) model [16]. It allows projects to support the storage and exchange of data from research partners based on the FAIR principles.

Research sharing offers the possibility to create ”snapshots” of an experiment. These snapshots can be upload in Zenodo or downloaded in an archive file.

Integration with the GenOuest infrastructure

The CeSGO environment is hosted on the GenOuest infrastructure and fully integrated with it. All services are linked to the same authentication system in order to avoid a new user registration.

A plugin for the GenOuest account manager was developed to automatically mount user's GenOuest directories into the Data-access service as external storage.
To properly monitor CeSGO usage and extract statistics, an in-house software was developed. Named CeSGO dashboard, this service allows GenOuest members to obtain information about the CeSGO environment

Keywords: Virtual Research Environment, collaboration, FAIR, data management, project management

BD_NGS, a complete tool for management, analysis and visualisation of NGS data for diagnosis purposes

Julien Plenecassagnes *† [1], Christophe Habib * ‡ [2], Nicolas Jeanne * [3], Frédéric Escudié [4], Laura Do Souto Ferreira [5], Manon Cassou * [1]

[1] Direction des Systèmes d'Information (DSI) – IUCT Oncopole, Institut Claudius Regaud – France
[2] Secteur de Biologie Moléculaire et Génétique Constitutionnelle, Plateau des Techniques Spécialisées (SBMGC-PTS) – CHU Toulouse – Institut Fédératif de Biologie, CHU Purpan, France
[3] Plateau Technique d'Infectiologie (PTI) – CHU Toulouse – Institut Fédératif de Biologie, CHU Purpan, France
[4] Laboratoire d'Anatomo-Cytopathologie – IUCT Oncopole – France
[5] Unité de Génomique du Myélome 4127, Equipe 13 CRCT (UGM 4127-Eq13) – IUCT Oncopole, CRCT toulouse – France

The Institut Universitaire du Cancer de Toulouse (IUCT) is a teaching hospital group of public and private partners, the Centre Hospitalier Universitaire (CHU) de Toulouse and the Institut Claudius Regaud (ICR), respectively. The IUCT encompass 39 regional hospitals and treats every year 80000 patients. We currently have 6 bioinformaticians within the IUCT: 4 are embedded in diagnosis labs of the CHU de Toulouse, and 2 are attached to the ICR whose work is shared by the whole IUCT.
In this abstract, we present our high performance computing (HPC) cluster dedicated to NGS analysis, hosting BD_NGS, an *in situ* developed tool.

The IUCT invested the HPC cluster v1 in 2015. The configuration has been set up by the society Axians with DELL and IBM hardware. This cluster has 8 computing nodes (including an interactive one) with 256Go of RAM each, a total of 165 To for storage/computing and 72 To, replicated for archive. We plan to increase the size of the existing cluster to be able to analyse data of a sequencer such as the Illumina NovaSeq 6000, to comply with the increase of activity. In parallel, the CHU de Toulouse plans to build a new team to propose a service of BigData analysis using this cluster.

BD_NGS is a GWT (Google Web Toolkit) [1] tool with a MariaDB [2] database for NGS results, which encompass a wide range of functionality. First, it automatically manages the life cycle of the data by retrieving the raw data from the sequencers, launching the analysis pipelines and

*Speaker

†Corresponding author: Plenecassagnes.Julien@iuct-oncopole.fr

‡Corresponding author: habib.c@chu-toulouse.fr

performing the archiving process. We currently write Snakemake pipelines [3], using Dockers containers [4] provided by the bioinformatics community and our own team within Singularity registry [5] to answer reproducibility issues. Pipelines are tested weekly for quality purpose. BD_NGS provides benefits at different levels. The technicians can create worksheets for validation purpose. The biologists, for diagnosis purpose, have access to a user-friendly interface of annotated NGS results and can interpret the variants and finally generate reports according to the ANPGM recommendations [6]. The ultimate goal would be to integrate BD_NGS with the IUCT laboratory information management system, in order to transmit its results. Throughout the process, raw data, run metrics and pipelines results are automatically archived.

BD_NGS is already used in production and is a central tool for the IUCT NGS analysis.

References:

1 - Web site: www.gwtproject.org

2 - Web site: https://mariadb.com

3 - Johannes Koster, Sven Rahmann; Snakemake-a scalable bioinformatics workflow engine, Bioinformatics, Volume 28, Issue 19, 1 October 2012, Pages 2520–2522, doi.org/10.1093/bioinformatics/bts480

4 - Web site: https://www.docker.com

5 - Kurtzer, Gregory M., Vanessa Sochat, et Michael W. Bauer. " Singularity: Scientific Containers for Mobility of Compute ". *PLOS ONE* 12, no 5 (11 mai 2017): e0177459. https://doi.org/10.1371/journ
6 - Web site: https://www.anpgm.fr

Keywords: NGS, Genomic, GWT, Snakemake, BD_NGS, IUCT, ICR, CHU de Toulouse, Diagnosis, HPC, ANPGM, Docker, Singularity

The journey of a team of engineers in learning packaging technology

Valentin Marcon *† 1, Laure Quintric‡ 2, Durand Patrick§ , Olivier Inizan¶ 1, Caroline Dussart 3, Valentin Loux 4, Maria Bernard 5, Géraldine Pascal 6

[1] Plateforme Migale, MaIAGE (MaIAGE) – Institut National de la Recherche Agronomique - INRA (FRANCE) – Centre de Jouy-en-Josas Domaine de Vilvert F78352 JOUY-EN-JOSAS Cedex, France
[2] Service Ressources Informatiques et Communications (IMN/IDM/RIC) – Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) – Centre de Brest, Pointe du Diable, F-29280 Plouzané, France, France
[3] Ifremer (Cellule de bioinformatique) – Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) – France
[4] Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE) – Institut national de la recherche agronomique (INRA) : UR1404 – Bâtiment 210-233 Domaine de Vilvert 78350 Jouy en Josas Cedex, France
[5] Système d'Information des GENomes des Animaux d'Elevage (SIGENAE) – Institut national de la recherche agronomique (INRA) : UMR1313 – France
[6] Institut National de Recherche Agronomique - Centre de Toulouse (INRA TOULOUSE) – Institut national de la recherche agronomique (INRA) : UMR1388 – Chemin de Borde Rouge BP52627 31326 Castanet Tolosan Cedex, France

Imagine that you have a wonderful pipeline that finds, rapidly, OTUs (operational taxonomic units). Imagine that this pipeline meets a growing success in the metagenomic community.

As engineers working on bioinformatics platforms you want to package it and distribute it. Despite the fact that this pipeline is already wrapped for Galaxy (yes, you can Find, Rapidly, OTUs with a Galaxy Solution aka FROGS [Escudie et al., Bioinformatics, 2017]), the point is that this pipeline is built upon on more than 20 dependencies. Again, you are an enthusiast engineer but you are far from being an expert of packaging and deployment.
In this poster we want to relate the journey of packaging and distributing a pipeline that enables a significant amount of dependencies. We are engineers in charge of setting up production quality bioinformatics services for the community of users of several French academic research institutes (IFREMER and INRA). On a daily basis, part of our work consists in installing and configuring softwares and we would like to have these steps as efficient as possible. At the beginning of the journey we had no particular skills in technologies for packaging of softwares and their deployment on a mutualized infrastructure.

We will focus on the learning curve for the technologies and tools employed (CONDA and

*Speaker
†Corresponding author: valentin.marcon@inra.fr
‡Corresponding author: laure.quintric@ifremer.fr
§Corresponding author: Patrick.Guido.Durand@ifremer.fr
¶Corresponding author: olivier.inizan@inra.fr

PLANEMO). We will show how (i) the fact that all team members decided to learn together and (ii) the support and reactivity of the developer community has speed up the learning process. We will also relate the early installation experiences done by people not involved in the packaging and with also no particular skills in dependencies resolution technologies.
Based on this experience, we will give a feedback on some easy-to-implement rules that would greatly simplify tasks to package, test and deploy complex bioinformatics pipelines right into Galaxy.

Keywords: Packaging, Galaxy, Metagenomic, CONDA, PLANEMO

IMGT/mAb-DB and IMGT/2Dstructure-DB for IMGT standard definition of an antibody: from receptor to amino acid changes

Mélissa Cambon *† 1, Karima Cherouali 1, Anjana Kushwaha 1, Véronique Giudicelli 1, Patrice Duroux‡ 1, Sofia Kossida§ 1, Marie-Paule Lefranc¶ 1

[1] IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM) – Institut de Génétique Humaine (IGH) CNRS Université de Montpellier UMR 9002 – 141 rue de la Cardonille 34396 Montpellier Cedex 5, France

INTRODUCTION

IMGT®, the international ImMunoGeneTics information system®, http://www.imgt.org [1], is the global reference in immunogenetics and immunoinformatics [2], founded in 1989 by Marie-Paule Lefranc at Montpellier (Université de Montpellier and CNRS). IMGT® is a high-quality integrated knowledge resource specialized in the immunoglobulins (IG) or antibodies, T cell receptors (TR), major histocompatibility (MH) of humans and other vertebrate species, and in the immunoglobulin superfamily (IgSF), MH superfamily (MhSF) and related proteins of the immune system (RPI) of vertebrates and invertebrates.

IMGT® has been built on the IMGT-ONTOLOGY axioms and concepts, which bridged the gap between genes, sequences, and three-dimensional (3D) structures. The concepts include the IMGT® standardized keywords (concepts of identification), IMGT® standardized labels (concepts of description), IMGT® standardized nomenclature (concepts of classification), IMGT unique numbering, and IMGT Colliers de Perles (concepts of numerotation) [2]. IMGT® comprises seven databases, 15,000 pages of web resources, and 17 online tools [1]. Annotated data of IMGT/mAb-DB [3] and IMGT/2Dstructure-DB [4] are being used to generate the IMGT standard definition of an antibody, from receptor to amino acid changes.

METHODOLOGY

Therapeutic proteins found in IMGT/mAb-DB and IMGT/2Dstructure-DB include the IG or antibodies (defined as containing at least one IG variable domain), the fusion protein for immune application (FPIA), the composite protein for clinical application (CPCA) and related protein of the immune system (RPI).

*Speaker

†Corresponding author: melissa.cambon@igh.cnrs.fr

‡Corresponding author: patrice.duroux@igh.cnrs.fr

§Corresponding author: sofia.kossida@igh.cnrs.fr

¶Corresponding author: marie-paule.lefranc@igh.cnrs.fr

IMGT/2Dstructure-DB on-line since 2001 contains 5056 entries of which 3531 are IG with amino acid (AA) sequences from different sources (2821 PDB, 374 INN, 336 Kabat). IMGT/2Dstructure-DB was implemented on the model of IMGT/3Dstructure-DB in order to manage AA sequences of multimeric receptors. Each chain is described in the "Chain details" section which comprises information first on the chain itself, then per domain. Chain and domain annotation includes the IMGT gene and allele names (CLASSIFICATION), region and domain delimitations (DESCRIPTION) and domain AA positions according to the IMGT unique numbering (NUMEROTATION). The closest IMGT® genes and alleles (found expressed in each domain of a chain) and the complementarity determining region (CDR)-IMGT lengths are identified with the integrated IMGT/DomainGapAlign tool [5], which aligns the AA sequences with the IMGT/DomainDisplay AA domain reference sequences [1]. The IMGT reference sequences are acquired by all the upstream work of manual biocuration.

IMGT/mAb-DB, the IMGT database created as an interface for therapeutic proteins, contains 797 entries which comprise 682 IG, 25 FPIA, 44 CPCA and 41 RPI. IMGT/mAb-DB provides the receptor identification in one of the categories (IG, FPIA, CPCA, RPI, and potentially TR and MH if entries become available), links to IMGT/2Dstructure-DB (for entries with AA sequences available) and to IMGT/3Dstructure-DB (for entries with three-dimensional structures available), target name with the HGNC nomenclature (and cross-reference to it), clinical indications, authority decisions and links related to them.

The IMGT standard definition of an antibody can be generated from the IMGT annotated data for the IG entered in both databases, whatever its format (complete IgG, Fab, F(ab')2, scFv...) and whatever its species (*Homo sapiens*, *Mus musculus*, chimeric, humanized...).

RESULTS

Data coming from IMGT/mAb-DB and IMGT/2Dstructure-DB involved in the composition of the sentences of the IMGT standard definition of an antibody includes: specificity, receptor identification, chain identification, positions of domains and disulfide bridges, mutations and the closest IMGT V and J genes and alleles of the amino acid sequences, CDR-IMGT lengths, amino acid changes (polymorphic or engineered).

Currently, the data model (both the Java classes in the implementation of IMGT/mAb-DB and the relational database schema using mapping technology) handles the receptor identification as a linear syntactical construction using terms (basic lexicon) and operators. Parenthesis or brackets grouping (for series-parallel, respectively) are shown with an optional number suffix (including 1 in specific case), and two combinators '-' and '_' for fusion and covalent association between different chains of a receptor. If the substance is a result of a fusion the suffix 'fusion' is added.

Moreover, IMGT/mAb-DB entries are illustrated by a graphical representation which is an "abstract picture" providing visualization of the different categories of the therapeutic substances.

Polymorphic AA changes of the allotypes are described with their position in the IMGT domain and chain and by their position in the sequence. Allotypes are allelic antigenic determinants identified in humans on the immunoglobulin (IG) gamma1, gamma2, gamma3 and alpha2 heavy chains (they are designated as G1m, G2m, G3m and A2m allotypes, respectively), and on the kappa light chain (Km allotypes) [6]. Recently, allotypes regained a lot of attention, owing to the development of therapeutic monoclonal antibodies and their potential immunogenicity. Therapeutic antibodies are most frequently of the IgG1 isotype, and to avoid a potential immunogenicity, the constant region of the gamma1 chains are often engineered to replace the

G1m3 allotype by the ”less immunogenic” G1m17 (CH1 R120> K) (G1m17 is more extensively found in different populations) [6]. The description links the allotypes to the IGHG and IGKC genes and alleles, respectively [2].

CONCLUSIONS AND PERSPECTIVES

The IMGT standard definition of an antibody is, per se, a paradigm for any other protein, receptor or ligand, natural, engineered or synthetic. Indeed, antibodies are widely used in clinical applications and for therapeutic purposes owing to their high level of specificity and affinity and to their structure in domains well fitted for antibody engineering and very diverse novel formats.

The main syntactical expressions are necessary and sufficient to provide a standardized IMGT definition schema for any antibody or related receptor type at the receptor, chain, domain and AA level.

They include the species, the IMGT receptor type with an optional complement for radiolabelled or conjugated or fused elements (identification) and a list of chain and domain labels (description) including identity percentage to the closest IMGT gene or allele (classification and numerotation).

Amino acids in the IGHG constant regions of the IG heavy chains are frequently engineered to modify the effector properties of the therapeutic monoclonal antibodies. In order to enrich the description with that information, we recently establish the IMGT engineered variant nomenclature [7] (using the IMGT unique numbering and IMGT chain and domain) for positions of the AA changes involved in antibody-dependent cellular (ADCC), antibody-dependent cellular phagocytosis (ADCP), complement-dependent cytotoxicity (CDC), half-life, reported in the literature. The IMGT engineered variant nomenclature [6] also includes AA changes at positions of interest in antibody engineering: knobs-into-holes AA changes is a rational design strategy, used for heterodimerization of the heavy (H) chains, in the production of bispecific IgG antibodies and controlling half-IG exchange is also a strategy for the generation of bispecific IgG1.

Implementation of Natural Language Processing (NLP) techniques in a Java tool will be considered in order to generate the definition as automatically and accurately as possible.

REFERENCES

Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, Hadi-Saljoqi S, Sasorith S, Lefranc G, Kossida S. IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucl. Acids Res., 43(Database issue):D413-22 (2015). PMID: 25378316

Lefranc M-P. Immunoglobulin (IG) and T cell receptor genes (TR): IMGT® and the birth and rise of immunoinformatics. Front Immunol., 5:22 (2014). PMID: 24600447

Poiron C, Wu Y, Ginestoux C, Ehrenmann F, Duroux P and Lefranc M-P. *IMGT/mAb-DB: the IMGT® database for therapeutic monoclonal antibodies. 11èmes Journéees Ouvertes de Biologie, Informatique et Mathématiques (* JOBIM), Montpellier, September 7-9, Abstract 13 (2010).

http://www.sfbi.fr/sites/default/files/jobim/jobim2010/index59385938.html?q=fr/node/55#IMGTmAb-DB:_the_IMGT_database_for_therapeutic_monoclonal_antibodies._

Kaas, Q., Ruiz, M. and Lefranc, M.-P. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucl. Acids. Res., 32:D208-D210 (2004). PMID: 14681396

Ehrenmann F, Kaas Q, Lefranc M-P. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. Nucl. Acids Res., 38:D301-307 (2010). PMID: 19900967

Lefranc M.-P, Lefranc G. Human Gm, Km and Am allotypes and their molecular characterization: a remarkable demonstration of polymorphism. In: B. Tait, F. Christiansen (Eds.), Immunogenetics, chap. 34, Humana Press, Springer, New York, USA. Methods Mol. Biol. 2012; 882, 635-680. PMID: 22665258

Cambon M, Sasorith S, Lefranc M-P. Amino acid positions involved in ADCC, ADCP, CDC, half-life and half-IG exchange. IMGT®, the international ImMunoGenetics information system® http://www.imgt.org.

http://www.imgt.org/IMGTeducation/Tutorials/IGandBcells/_UK/IGproperties/Tableau3.html

Created: 11/12/2012. Version: 12/12/2017

Keywords: IMGT, immunoinformatics, immunogenetics, immunoglobulin, antibody, IMGT/mAb, DB, IMGT/2Dstructure, DB

IMGT/HighV-QUEST statistical analysis of IMGT clonotypes (AA), novel interface and functionalities for NGS analysis of IG and TR

Marianne Lèbre [*][†] [1], Karthik Kalyan[‡] [1], Patrice Duroux[§] [1], Véronique Giudicelli[¶] [1], Sofia Kossida[‖] [1], Marie-Paule Lefranc[**] [1]

[1] IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM) – Institut de Génétique Humaine (IGH) CNRS Université de Montpellier UMR 9002 – 141 rue de la Cardonille 34396 Montpellier Cedex 5, France

Introduction

IMGT®, the international ImMunoGeneTics information system®, http://imgt.org/ [1], is the global reference in immunogenetics and immunoinformatics [2], founded in 1989 by Marie-Paule Lefranc at Montpellier (Université de Montpellier and CNRS). IMGT® is a high-quality integrated knowledge resource specialized in the immunoglobulins (IG) or antibodies, T cell receptors (TR), major histocompatibility (MH) of humans and other vertebrate species, and in the immunoglobulin superfamily (IgSF), MH superfamily (MhSF) and related proteins of the immune system (RPI) of vertebrates and invertebrates.

IG and TR are the antigen receptors of the adaptive immune response which characterizes the vertebrates with jaws (*gnasthostomata*) [2]. Their study in normal and pathological conditions is a challenge due to their huge diversity (1012 potential specificities for humans) only limited by the number of the B and T cells that an organism is genetically programmed to produce. The high diversity of the variable domain at the N-terminal end of each IG or TR chain results from genomic DNA rearrangements which occur in B or T cells, respectively, and which involve variable (V), diversity (D) and joining (J) genes [2]. This combinatorial V-(D)-J diversity is further increased by the junctional diversity and for IG, by somatic hypermutations. The analysis of the immune repertoires has become feasible thanks to the developments of the next generation sequencing (NGS) technologies. Since 2010, IMGT® has developed IMGT/HighV-QUEST [3, 4], the high-throughput version of IMGT/V-QUEST, which is so far the only online tool available on the Web for the direct analysis of complete IG and TR variable domains (V-DOMAIN, corresponding to the V-(D)-J REGION) of NGS nucleotide rearranged sequences, from humans and other vertebrate species.

[*] Speaker

[†] Corresponding author: marianne.lebre@igh.cnrs.fr

[‡] Corresponding author: karthik.kalyan@igh.cnrs.fr

[§] Corresponding author: Patrice.Duroux@igh.cnrs.fr

[¶] Corresponding author: veronique.giudicelli@igh.cnrs.fr

[‖] Corresponding author: sofia.kossida@igh.cnrs.fr

[**] Corresponding author: Marie-Paule.Lefranc@igh.cnrs.fr

METHODOLOGY

IMGT/HighV-QUEST analyzes up to 500,000 sequences per run, with the same degree of resolution and high-quality results as IMGT/V-QUEST and IMGT/JunctionAnalysis [3, 4]. Indeed IMGT/HighV-QUEST uses the same algorithm and runs against the same IMGT reference directories. IMGT/HighV-QUEST numbers the user sequences according to the IMGT unique numbering and introduces gaps accordingly. It identifies the V, D and J genes in rearranged IG and TR sequences and, for the IG, characterizes the nt mutations and amino acid (AA) changes resulting from somatic hypermutations by comparison with the IMGT/V-QUEST reference directories. The tool integrates IMGT/JunctionAnalysis for the detailed characterization of the V-D-J or V-J junctions, IMGT/Automat for a complete sequence annotation with the delimitation of the IMGT labels of description. By default, IMGT/HighV-QUEST identifies the insertions/deletions (indels) which are NGS errors resulting from homopolymer hybridization and corrects them. IMGT/HighV-QUEST results consists classically of 11 CSV text files downloadable as an archive file [3, 4]. The CSV files contain one line per analysed sequence, and together may comprise up to 539 columns for a complete results report.

The IMGT/HighV-QUEST statistical analysis, which allows the identification and characterization of the clonotypes [5], may analyse up to one million IMGT/HighV-QUEST results.

RESULTS

In the literature, clonotypes characterize the repertoires of the adaptive immune responses but are defined differently, depending on the experiment design (functional specificity) or available data. Thus, a clonotype may denote either a complete receptor (e.g., TR-alpha_beta), or only one of the two chains of the receptor (e.g., TRA or TRB), or one domain (e.g., V-BETA), or the CDR3 sequence of a domain. Moreover the sequence can be at the AA or nt level, and this is rarely specified. Therefore, IMGT priority was to define clonotypes and their properties, which could be identified and characterized by IMGT/HighV-QUEST within the statistical analysis, unambiguously. In IMGT, the clonotype, designated as 'IMGT clonotype (AA)', is defined by a unique V-(D)-J rearrangement (with IMGT gene and allele names determined by IMGT/HighV-QUEST at the nt level) and a unique CDR3-IMGT AA junction sequence [5]. An IMGT clonotype (nt) is defined by a unique V-(D)-J rearrangement (with IMGT gene and allele names determined by IMGT/HighV-QUEST at the nt level) and a unique CDR3-IMGT nt junction sequence. Several IMGT clonotypes nt may correspond to one IMGT clonotype (AA).

The statistical analysis applies a filter on the IMGT/HighV-QUEST results: only the ones characterized by a V-GENE and allele (single or several alleles), a JUNCTION and a J-GENE and allele (single or several alleles) are filtered-in for statistical analysis [5]. Statistical analysis output is provided as a txz file (IMGT/HighV-QUEST Documentation: http://www.imgt.org/HighV-QUEST/doc.action). In order to evaluate and to explore, between sets, the significance of pairwise comparison of IMGT clonotype (AA) diversity and expression per V, D and J gene, IMGT/StatClonotype [6] was developed. This tool is downloadable on the IMGT® site (http://www.imgt.org/S Integrated in the R package "IMGTStatClonotype", it offers a graphical interface to visualize pair wise comparison, per IMGT genes and alleles, of the IMGT clonotype (AA) diversity or expression of any IG or TR immunoprofiles of any species, obtained as outputs of the IMGT/HighV-QUEST statistical analysis.

With the advent of single molecule, long read sequencing (PacBio), the advanced functionality "Analysis of single chain Fragment Variable (scFv)" sequences [7] has been added as an option within IMGT/HighV-QUEST. So far, the NGS analysis of scFv was a challenge. Indeed, scFv are engineered antibody single chain fragments which comprise two variable domains associated by a peptide linker and the NGS methods did not provide reads long enough to span the length of the scFv (> 800 bp). If selected, the IMGT/HighV-QUEST functionality analyses both V-DOMAIN individually (results in the 11 CSV files) and produces a 12th CSV result file "scFv" where the association between the two V-DOMAIN, their respective positions in the sequence, the positions and the length of the linker are recorded. This advanced functionality allows the analysis of the content of scFv combinatorial phage display libraries which are classically screened for identification of novel therapeutic antibody specificities.

CONCLUSIONS AND PERSPECTIVES

IMGT/HighV-QUEST is the standard for the NGS analysis of IG and TR repertoires in experimental engineered (combinatorial libraries) or in physiological conditions (vaccination, immunodeficiency, autoimmune diseases, cancers and infectious diseases). IMGT/HighV-QUEST is particularly well adapted for the analysis of complete V domains of the IG and TR repertoires from B and T subsets, in many experiments and from many individuals (humans or other vertebrate species).

IMGT/HighV-QUEST was originally developed using Java based technologies and was initially a 2-tier architecture system: application, database. It combined a web-based user interface (client UI) and a job management system (using Java Quartz) in one web application. It evolved into a 3-tier architecture system: client UI, database and scheduling-system. The scheduling-system is now a standalone system (shell scripts and cron) which has the possibility to be integrated to an automation tool, such as Rundeck. After the implementation of the 3-tier architecture system, a new client UI will soon be made available based on modern web technologies (Bootstrap, Struts2 and Tiles3). The 3-tier architecture will enable easier implementation of the newly developed functionalities.

IMGT/HighV-QUEST makes use of High Performance Computing (HPC) clusters to run large number of user submitted jobs that are split into tasks (i.e. IMGT/V-QUEST runs) based on a linear equation. In regards to the new available HPC clusters and the users' demands to increase the number of sequences in one job, the reduction in computational processing time by parallelizing the IMGT/V-QUEST module within IMGT/HighV-QUEST is underway with the usage of dynamic parameters. Java 'multi-threading (fork-join/work-stealing-algorithm) and profiling benchmarking (JMH APIs)' are used for its development. The possibility of utilizing distributed computing facilities (JPPF) is also explored.

ACCESS TO HPC RESOURCES

This work was granted access to the HPC@LR and to the HPC resources of the Centre Informatique National de l'Enseignement Supérieur (CINES) and to Très Grand Centre de Calcul (TGCC) of the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) under the allocation [036029] (2010-2018) made by GENCI (Grand Equipement National de Calcul Intensif).

REFERENCES

Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, Hadi-Saljoqi S, Sasorith S, Lefranc G, Kossida S. IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucl. Acids Res. 2015 Jan;43(Database issue):D413-22. doi: 10.1093/nar/gku1056. Epub 2014 Nov 5 Free article. PMID: 25378316

Lefranc M-P. Immunoglobulin (IG) and T cell receptor genes (TR): IMGT® and the birth and rise of immunoinformatics. Front Immunol. 2014 Feb 05;5:22. doi: 10.3389/fimmu.2014.00022. Open access. PMID: 24600447

Alamyar E., Giudicelli V. Duroux P, Lefranc, M.-P. *IMGT/HighV-QUEST: a high-throughput system and web portal for the analysis of rearranged nucleotide sequences of antigen receptors - High-throughput version de IMGT/V-QUEST*. Poster n◦27 (abstract n◦60). 11èmes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM 2010).Montpellier, France (7-9 septembre 2010).

http://www.sfbi.fr/sites/default/files/jobim/jobim2010/index59385938.html?q=fr/node/55#IMGTHighV-QUEST:_A_High-Throughput_System_and_Web_Portal_for_the_Analysis_of_Rearranged_Nucleotide_Sequences_c _High-Throughput_Version_of_IMGTV-QUEST_

Alamyar E, Giudicelli V, Shuo L, Duroux P, Lefranc M-P. IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Res. 2012, April 20;8:1:2. doi: 10.4172/1745-7580.1000056. PMID: 22647994

Li S., Lefranc M.-P., Miles J.J., Alamyar E., Giudicelli V., Duroux P., Freeman J.D., Corbin V.D.A., Scheerlinck J.-P., Frohman M.A., Cameron P.U., Plebanski M., Loveland B., Burrows S.R., Papenfuss A.T., Gowans E.J. IMGT/HighV-QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat. Commun. 2013;4:2333. doi:10.1038/ncomms3333 Open access. PMID: 23995877

Aouinti S, Giudicelli V, Duroux P, Malouche D, Kossida S, Lefranc M-P. IMGT/StatClonotype for Pairwise Evaluation and Visualization of NGS IG and TR IMGT Clonotype (AA) Diversity or Expression from IMGT/HighV-QUEST. Front Immunol. 2016 Sep 9;7:339. doi: 10.3389/fimmu.2016.00339. eCollection 2016. Free PMC Article. PMID: 27667992

Giudicelli V, Duroux P, Kossida S, Lefranc M-P. IG and TR single chain Fragment variable (scFv) sequence analysis: a new advanced functionality of IMGT/V-QUEST and IMGT/HighV-QUEST. BMC Immunol. 2017 Jun 26;18(1):35. doi: 10.1186/s12865-017-0218-8. PMID: 28651553.

Keywords: IMGT, immunoinformatics, immunogenetics, immunoglobulin, antibody, receptor T, IMGT clonotype (AA), NGS

JASS: visualising the results of a joint analysis in the context of GWAS

Pierre Lechat * [1], Herve Menager [1], Vincent Guillemot [1], Hanna Julienne [1], Hugues Aschard [1], Vincent Laville [1], Bjarni Vilhjalmsson [2]

[1] Institut Pasteur [Paris] – - – 25-28, rue du docteur Roux, 75724 Paris cedex 15, France
[2] Department of Computer Science [Aarhus] – Department of Computer Science Aarhus University Åbogade 34 DK-8200 Aarhus N Denmark, Denmark

Thousands of Genome Wide Association Studies (GWAS) have been performed in human cohorts, identifying numbers of genetic loci associated with human traits and diseases. Moving forward, recent works highlighted that summary statistics from these GWAS can be used to perform joint analyses of multiple phenotypes , potentially allowing for enhance statistical power and therefore, the detection of new associations missed by univariate GWAS. However, in practice, billions of possibles combinations can be builds from a few dozen GWAS, while interpretation of joint test often requires a comparison with individual GWAS association signals. It follows that the field can strongly benefit from a user-friendly and efficient tool to perform such analyses. .
Here, we present JASS (Joint Analysis of Summary Statistics), a software tool providing an interactive web interface that addresses that need. JASS allows for a very fast application of joint test across dozens of GWAS and an interface that provides a range of results' visualisations at different genomic scales. In particular, JASS uses the javascript library plotly.js in order to build dynamic heatmaps representing the individual GWAS results, synchronized with a dynamic Manhattan plot of the joint analysis results. The duality between global and local visualisations provides the user with a satisfying level of interaction with the analyses while being able to manage a fair amount of data (more than 1M SNPs for 45 analyses). In a nutshell, JASS answers to the need of exploring a lot of data characterized by a very sparse signal.

A version of this application is available online at
http://jass.pasteur.fr. JASS is freely available at https://gitlab.pasteur.fr/statistical-genetics/jass. It can be installed locally and run either as a web server or command line tool for advanced users.

Keywords: GWAS, joint analysis, visualisation

[*] Speaker

New selected set of species for the Quest for Orthologs Consortium

Alan Sousa Da Silva * [1]

[1] European Bioinformatics Institute [Hinxton] (EMBL-EBI) – EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK, United Kingdom

Since the first QfO data set release in 2010, UniProt has been providing data for 66 reference species, community chosen during the first QfO meeting (2009). These species formed the basis for the UniProt Reference Proteomes project [1], which has subsequently been extended to a larger number of species to entail a much broader sequenced taxonomy hierarchy [2].

The UniProt Reference Proteomes currently comprises more than 9,000 species, 2,245 of which are viruses, most not within the scope of QfO. However, there is a subset of species available for the QfO community composed of 245 manually curated species [3]. This reduced set of species give us a reasonable coverage of the Tree of Life and hence a new opportunity to review and revise the original list of species used for QfO standardised benchmarks. To help with this, we have made use of the Proteome Priority Score (PPS) [4], a more robust and reliable way of prioritising proteomes based on completeness and the quality of their annotations. PPS compiles the number of proteins, publications and functional annotation scores for a given proteome, thus assisting us to sort and select one species that better represents all the species within the phylum.

This approach for the selection of new species within the QfO initiative allowed us to extend the data set that now comprises 78 species (7 Archaea, 23 Bacteria and 48 Eukaryota), where 16 new species were added and 4 removed. The new selected species set will be released in April 2018 and will contains FASTA files for the proteins and DNA sequences, mapping files (UniProt accessions to others cross-referenced databases) and a file in SeqXML format encompassing all data provided.

UniProt endeavours to support biological research by providing a stable, comprehensive, consistent and accurately annotated protein knowledgebase that is freely accessible for the scientific community. In addition to standardising and integrating data from numerous resources, UniProt provides rich and comprehensive functional annotation of its protein sequences.

UniProt Knowledgebase (UniProKB). The UniProt Consortium: Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Research (2012). DOI: 10.1093/nar/gkr981

ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes

ftp.ebi.ac.uk/pub/databases/reference_proteomes/QfO

[*] Speaker

Chen C, Natale DA, Finn RD, Huang H, Zhang J, et al. Representative Proteomes: A Stable, Scalable and Unbiased Proteome Set for Sequence Analysis and Functional Annotation. Plos One (2011). DOI: 10.1371/journal.pone.0018910

Keywords: reference proteomes, PPS, orthologs, seqxml, uniprot

AuBi platform for biologists and bioinformaticians at UCA Mesocentre

Nadia Goué * [1], David Grimbichler , Antoine Mahul

[1] Plateforme AuBI - Mésocentre - Université de Clermont-Ferrand (UCA) – Direction des Systèmes d'Information – 7, avenue Blaise Pascal CS 60026 63 178 Aubière cedex, France

The Mesocentre as part of Clermont Auvergne University (UCA) is delivering services in sciences data computing (HPC, VM, ...) and short-term storage through a network of technology core facilities. These offers are done to assist multi-disciplinary scientists in their computing projects. At that time, we are hosting a computer farm with about 800 cores, 40 nodes for moderate memory usage (< 256 Gb) and a SMP supercomputer made of 384 cores and 12 To memory in addition to a CEPH storage of at least 1 To capacity by user.
Hosted by the Mesocentre, the Auvergne bioinformatics (AuBi) platform is a member of the French Bioinformatics Institute (IFB, https://www.france-bioinformatique.fr/en/platforms/AUBI). AuBi platform aims at sharing expertises and knowledge in large-scale data treatments and analysis by supplying a complete computing environment with hardware and software infrastructures for 9 research laboratories. AuBi platform is then involved in various projects belonging to genomics, metagenomics, transcriptomics, modeling and imaging fields amongst others [1,2,3]. Furthermore, we provide support to UCA laboratories and Associates in their effort to maintain and enhance their scripts and pipelines used on our infrastructure.

Another aspect of AuBi platform work is to facilitate computing access to non-bioinformatician biologists by the way of a Galaxy server released in the upcoming weeks. We are also organizing training sessions to help our users, either biologists or bioinformaticians to optimize computing resources usage through command line interface and Galaxy environment.

References

1. Amato P., Joly M., Besaury L. Oudart A., Taib N., Moné A., Deguillaume L., Delort A.M. and Debroas D. (2017). Active microorganisms thrive among extremely diverse communities in cloud water. PLoS ONE 12(8):e0182869.

2. Gasc C, Constantin A, Jaziri F, Peyret P: OCaPPI-Db: an oligonucleotide probe database for pathogen identification through hybridization capture. Database (Oxford) 2017, 2017.
3. Parisot N, Peyretaillade E, Dugat-Bony E, Denonfoux J, Mahul A, Peyret P: Probe Design Strategies for Oligonucleotide Microarrays. Methods Mol Biol 2016, 1368:67-82.

Keywords: Mesocentre, Ressources, Infrastructures, HPC, Storage

*Speaker

Taking tools out of their laboratory: methods and practices to make bioinformatics tools accessible.

Valentin Marcon *† [1], Alexis Dereeper [2], Sarah Maman [3], Luc Jouneau [4], Mélanie Petera [5], Marie Tremblay-Franco [6], Olivier Inizan‡ [7]

[1] MaIAGE (MaIAGE) – Institut National de la Recherche Agronomique - INRA (FRANCE) – Centre de Jouy-en-Josas Domaine de Vilvert F78352 JOUY-EN-JOSAS Cedex, France
[2] CIRAD UMR AGAP (AGAP) – Institut national de la recherche agronomique (INRA) : UMR1334 – TA A-108/03-Avenue Agropolis, 34398 Montpellier Cedex 5, France
[3] Institut National de Recherche Agronomique - Centre de Toulouse (INRA TOULOUSE) – Institut national de la recherche agronomique (INRA) : UMR1388 – Chemin de Borde Rouge BP52627 31326 Castanet Tolosan Cedex, France
[4] UMR BDR, INRA, ENVA, Université Paris Saclay, 78350, Jouy en Josas, France – Institut national de la recherche agronomique (INRA) : UMR1198 – Domaine de Vilvert, 78350 Jouy en Josas, France
[5] Université Clermont Auvergne, INRA, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont – Institut national de la recherche agronomique (INRA) : UMR1019 – F-63000 Clermont-Ferrand, France
[6] INRA Toxalim Axiom MetaToul MetaboHUB - Université de Toulouse – Institut National de la Recherche Agronomique - INRA : UMR1331, Institut national de la recherche agronomique (INRA) : UMR1331 – 180 chemin de Tournefeuille 31027 Toulouse, France
[7] Plateforme Migale, MaIAGE (MaIAGE) – Institut National de la Recherche Agronomique - INRA (FRANCE) – Centre de Jouy-en-Josas Domaine de Vilvert F78352 JOUY-EN-JOSAS Cedex, France

The tools accessibility is a well known story in the bioinformatics community. Often, when a tools is designed in a laboratory, considerations such as distribution and user-friendliness use are not well taken into account.

Several projects aim to make the tools more accessible. The Galaxy project for example, first focused on the user friendly use with a web framework, unifying the graphical user interface for all tools through a wrapping system. As the project grew, developers realized that not only user-friendliness must be taken into account for accessibility, but also packaging and installation, especially for tools with many dependencies.

Today, the Galaxy developer community is able to provide several methods and practices to make a tool accessible both to the user who's not comfortable with the command line and to the system administrator in charge of installing the tools.

This poster will present the result of the "Galaxy For Life Science" project (GFLS). This project was designed by the IFB Galaxy working group as a development resource for partner laboratories, to improve the accessibility of their tools, with the methods and practices of the Galaxy community, in order to reach a wider community.

The raw material for the project consists of tools from different scientific communities (grouped

*Speaker

†Corresponding author: valentin.marcon@inra.fr

‡Corresponding author: olivier.inizan@inra.fr

in several use cases: plant science, statistical analysis, livestock, bacteria). Some of them were accessible (via a galaxy server) but none were fully distributable.
We will show how we have used the methods and practices used and promoted by the galaxy developer community, especially the technologies used most recently (Conda, Containers), to make tools accessible both for the end users and also for the system administrators.

Keywords: Accessibility, Galaxy, Package, Conda, Container, Pipeline, Workflow

Bioinformatics services provided at Institut Curie

Marc Deloger *† [1,2,3], Elodie Girard [1,2,3], Maude Ardin [1,2,3], Dimitri Desvillechabrol [1,2,3], Isabel Brito [1,2,3], Pierre Gestraud [1,2,3], Stéphane Liva [1,2,3], Alexandre Sta [1,2,3], Laetitia Chanas [1,2,3], Choumouss Kamoun [1,2,3], Georges Lucotte [1,2,3], Henri De Soyres [1,2,3], Julien Romejon [1,2,3], Frédéric Jarlier [1,2,3], Philippe La Rosa [1,2,3], Patrick Poullet [1,2,3], Nicolas Servant [1,2,3], Philippe Hupe [1,2,3,4], Emmanuel Barillot [1,2,3]

[1] Institut Curie, PSL Research University, F-75005 Paris, France – Institut Curie, PSL Research University – France
[2] INSERM, U900, F-75005 Paris, France – Inserm – France
[3] MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France – MINES ParisTech, PSL Research University – France
[4] CNRS, UMR144, F-75005 Paris, France – CNRS : UMR144 – France

The Institut Curie Bioinformatics platform is composed of bioinformaticians, biostatisticians and software engineers who offer a multidisciplinary expertise to support the biotechnological platforms, the research units and the hospital in their daily activities. Our skills range from statistical data analysis, data management, software development and high performance computing. We have five main missions: (1) collaborative support to biologists or clinicians for bioinformatics and biostatistics data analysis in research but also for NGS-based clinical diagnostic, (2) delivering advices and training in biostatistics and bioinformatics, (3) knowledge and data integration, (4) support to high performance computing and (5) coordination of bioinformatics activities within Institut Curie.

The Bioinformatics platform provides biologists and clinicians with a collaborative bioinformatics and biostatistics support for data analysis, covering all aspects of large scale data analysis (proteomics, RNA-seq, ChIP-seq, Exome, Whole Genome, etc.) and also analysis of low-throughput data (affymetrix/agilent microarrays). This covers experimental design, data modeling and statistical analysis, developing tools for multi-level data analysis whenever needed and contribution to scientific communication.

A large portfolio of trainings in statistics, bioinformatics and high performance computing is proposed in order to promote autonomous data analysis by biologists and clinicians as often as possible.

The Bioinformatics platform also integrates the data generated by the Institut Curie's Biotechnology platforms: genome, transcriptome or proteome array platforms, mass spectrometry, and next-generation sequencing (Illumina, PacBio). The ultimate goal of this data integration is to promote data and knowledge sharing across the institute by providing a seamless integration of all levels of information. Data integration covers LIMS management (and development if needed), development and setup of automatic pipelines for high-throughput data analysis (for

*Speaker

†Corresponding author: marc.deloger@curie.fr

data preprocessing, quality control and first level analysis), development and management of databases and user interfaces, dataflow management. Some pipelines are routinely used for diagnosis and precision medicine at the hospital.
The IC has a computing cluster with 2000 cores, 14 TB of RAM and 100TB of local scratch. The storage capacitty is 2 PB (mirrored data).

Keywords: RNAseq, CLIPseq, ChIPseq, ATACseq, TGS, WES, Precision Medicine, WGS, HiCseq, 4Cseq, RPPA, DIA, SWATH, MS, Long Reads, Single cell, CRISPR screening, Biostatistics, Microarrays, Data Integration, HPC, Trainings

The bio.tools portal of bioinformatics tools and services

Piotr Chmura , Kenzo-Hugo Hillion *† 1, Hervé Ménager , Jon Ison

1 Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI) – Institut Pasteur de Paris – France

During the last decades, computer science rapidly became a major actor in the field of biological research. One consequence is the massive production of software tools and services by a great number of teams and units leading to new challenges for the bioinformatics community. One of these challenges is to make these tools and services findable and discoverable. The most common ways for publishing a new software are research articles and communications at scientific conferences. These two options provide the opportunity to present this work to the desired community. However, trying to find the right bioinformatics tool to perform a given task remains a challenge, and there is a need for some gateways to query all available resources for a given field or topic. Several existing initiatives tackle this issue such as Biocatalogue [1], Debian Med [2] or Omictools [3].

The bio.tools registry [4, 5], which is part of the ELIXIR infrastructure, is a registry of bioinformatics tools and services. In this registry, each entry is described with detailed metadata that include a wide range of scientific, technical and administrative properties. The scientific description of the tools is standardized by using most notably the EDAM ontology [6] to describe the domain(s) of application and the functions performed in a machine-readable format (what kind of input is consumed, what the resource does, and what kind of file comes out of it). This rich description makes it possible to access a wide range of information about the resources, including for instance research article references, author information, license, etc.

Bio.tools is available as a web portal, as well as through a REST API, both of these allowing to either query the existing resources or to create new ones. Two projects, respectively named ReGaTe [7] and ToolDog [8], have been developed on top of this access point. The first one automates the registration of Galaxy [9] services published by a given server, through the automated mapping of the metadata stored in the Galaxy tool descriptions themselves. The second one helps generating tool descriptions for Galaxy or CWL [10], based on the information stored in the registry entries.

Bio.tools, available since early 2015, now includes over 180,000 annotations on some 10,000 resources. Access to this portal is unrestricted, contributions are open and welcome, and the registry content is available under open Creative Commons Attribution licence (CC BY 4.0).

References:

[*] Speaker

[†] Corresponding author: kehillio@pasteur.fr

Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., Lopez, R., Goble, C.A.: BioCatalogue: a universal catalogue of web services for the life sciences, Nucl. Acids Res., 2010. doi:10.1093/nar/gkq394

S. M´oller, H. N. Krabbenh´oft, A. Tille, D. Paleino, A. Williams, K. Wolstencroft, C. Goble, R. Holland, D. Belhachemi, C. Plessy (2010) ”Community-driven computational biology with Debian Linux” BMC Bioinformatics, 11(Suppl 12):S5

Henry VJ, Bandrowski AE, Pepin A-S, Gonzalez BJ, Desfeux A. OMICtools: an informative directory for multi-omic data analysis. Database: The Journal of Biological Databases and Curation. 2014;2014:bau069. doi:10.1093/database/bau069.

https://bio.tools

Jon Ison et al. Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Research, 44(D1):D38–D47, January 2016. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv1116.

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S. and Rice, P. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics, 29(10): 1325-1332. doi: 10.1093/bioinformatics/btt113

Olivia Doppelt-Azeroual, Fabien Mareuil, Eric Deveaud, Matúš Kalaš, Nicola Soranzo, Marius van den Beek, Bj´orn Gr´uning, Jon Ison, Hervé Ménager; ReGaTE: Registration of Galaxy Tools in Elixir, GigaScience, Volume 6, Issue 6, 1 June 2017, Pages 1–4, doi: 10.1093/gigascience/gix022

Hillion KH, Kuzmin I, Khodak A et al. Using bio.tools to generate and annotate workbench tool descriptions [version 1; referees: 4 approved]. F1000Research 2017, 6(ELIXIR):2074, doi: 10.12688/f1000research.12974.1

Enis Afgan et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research (2016) doi: 10.1093/nar/gkw343

Amstutz, Peter et al. (2016): Common Workflow Language, v1.0. Figshare. doi: 10.6084/m9.figshare.3115156.v
Retrieved: 15 37, Mar 09, 2017 (GMT)

Keywords: registry, ontology, elixir, tools, services

Framboisine mini-clusters for Southern countries

Ndomassi Tando * [1], François Sabot * † [2], Christine Dubreuil-Tranchant‡ [3]

[1] Institut de Recherche pour le Développement (IRD [France-Sud]) – Institut de recherche pour le développement [IRD] – 911 avenue Agropolis,BP 6450134394 Montpellier cedex 5, France
[2] Diversité, adaptation, développement des plantes (DIADE) – Institut de recherche pour le développement [IRD] : UR232, Université Montpellier II - Sciences et techniques – Centre IRD de Montpellier 911 av Agropolis BP 604501 34394 Montpellier cedex 5, France
[3] UMR DIADE – Institut de Recherche pour le Développement – France

Framboisine mini-clusters provides a solution to the Southern Countries in terms of high performance parallel computing trainings and prototyping, especially in bioinformatics. Indeed, these computing clusters, composed of Raspberry Pi small cards, do not need big air conditioning systems, high electrical power or extra building. A typical cluster is composed of several computer blades (length 60 cm x width 42 cm x height 5 to 10 cm), in need of a minimal redundant 800W electrical power, with a performing air conditioning system, in a specific computer room. Each of these blades generally costs between 2000 and 4000 euros. These clusters are very effective but their initial cost, the support and their use restrict their access to big universities or research institutes usually based in North. Another possible alternative solution is the "Cloud Computing". With the "Cloud Computing", resources provided by companies such as Amazon, Google or Microsoft can be used via the web. The user can go beyond the constraints of installation and infrastructure costs (at least part of it). Nevertheless, in that case, other issues can appear such as the amount of costs over the entire subscription contract period, the network access, a sufficient bandwidth but also confidentiality and the intellectual property rights of data (extended Patriot Act; Nagoya Protocol).

To develop the concept of Framboisine mini-clusters, we use the Standford "proof-of-concept" on a Raspberry Pi cluster with a application to the bioinformatic.

Framboisine mini-clusters cost around 2000 for a 16 calculation cores and 2TB redundant storage system. They can be installed in any kind of rooms not only in computer room and can be cooled with a simple fan. Therefore, the entire electric consumption of the system does not exceed 300W. The calculation of this system is of course less performing then a real cluster, but it allows to quickly train the users, to work on limited data and to prototype analyses before applying them on wider scale system. 3 Framboisine clusters are currently deployed to Southern partners (AfricaRice, Ivory Coast; FAST/Dassa, Benin; ISRA/LMI LAPSE, Senegal).

*Speaker

†Corresponding author: francois.sabot@ird.fr

‡Corresponding author: christine.tranchant@ird.fr

The deployed Framboisine mini-clusters use 3 different types of cards:

Raspberry B: (512 MB of RAM, 1 Core): OS Rasbian

Olinuxino : (2GB of RAM, 2 cores): OS debian dedicated

Cubietruck (1GB of RAM, 2 cores): OS Cubian

Each mini-cluster is composed of a master node and several nodes composed of different types of cards(Raspberry B, Olinuxino or Cubietruck).

Each type of cards has its own operating system contained in a sd card. We have developed operating system images containing the most common bioinformatic softwares and system configurations for each of these 3 types of cards. This solution allows users to add as many nodes as they want on a mini-cluster as long as they modify the configuration of the master node to add them.

After the study of several cards available, we have chosen to use a cubietruck card as master node for each cluster because it was the most powerful card at that time and the Cubian operating system was the most developed. The power supply is provided by usb hubs and nodes are linked together thanks to a 16 ports switch. The 2TB storage is contained in a Nas Synology that can be access through a NFS connection.

Teaching is the first concrete application for these mini-clusters in Southern as in Northern countries. Indeed, this kind of low cost cluster allows newbies to have access to resources usually only used by well trained employees because of the infrastructure costs.Then, launching a bad jobs on a several thousands euros machine and damaged it can be serious, while switching a node card on a mini-cluster Framboisine costs less than 50. Moreover, for security reasons, computing cluster are rarely accessible in a anonymous way, thus training several users on several sessions can be complicated. With a portable system such as the framboisine mini-cluster, it is not a problem anymore.

The second application of the mini-clusters is to give access to computing tools to southern partners,admittedly limited, but needing a low investment. Indeed, create a computing cluster requires around 10000 of machines (servers, air conditioning), not counting a computer room. A Framboisine cluster only cost from 2000 to 2500 and does not need additional adjustments, its compact design being a little bit more bigger than a ordinary personal computer.

This solution can be declined with all kinds of cards and can evolve with the evolution of the cards. Currently, we are working on a new type of minicluster composed of Raspberry Pi 3 nodes. In the future, we can imagine to use GPU cards in our next mini-clusters. To make the installation and transportation of the mini-cluster easier, we are currently working on plug and play packages of 2 types:

- A portable one in collaboration with a FabLab
- One designed to stay in a fixed location in collaboration with a french firm

Keywords: bioinformatics, computing, HPC, clusters, Southern, low cost, education, trainings

ProteoRE, a Galaxy-based infrastructure for annotating and interpreting proteomics data

Florence Combes * [1], Lien Nguyen [2], Maud Lacombe [3], Lisa Perus [4], Virginie Brun [5], Valentin Loux [6], Yves Vandenbrouck [7]

[1] Exploring the Dynamics of Proteomes (EDyP), BGE/U1038, INSERM/CEA/Université Grenoble Alpes (CEA/DRF/BIG/BGE/EDYP) – INSERM U1038, Université Grenoble Alpes – 17 rue des Martyrs 38054 Grenoble Cedex 9, France

[2] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France et EDyP (CEA) (INRA,CEA, INSERM, UGA) – Université de Grenoble, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble, Centre National de la Recherche Scientifique : FR3425, Institut National de la Santé et de la Recherche Médicale - INSERM : U1038 – France

[3] CEA Grenoble (BIG, Biologie à grande Echelle, EDyP) – INSERM U1038, Université Grenoble Alpes – 17 rue des Martyrs 38054 Grenoble Cedex 9, France

[4] Exploring the Dynamics of Proteomes (EDyP), BGE/U1038, INSERM/CEA/Université Grenoble Alpes (BIG) – Inserm : U1038, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble, Université Joseph Fourier - Grenoble I – 17 rue des Martyrs, F-38054 Grenoble, France

[5] Exploring the Dynamics of Proteomes (EDyP), BGE/U1038, INSERM/CEA/Université Grenoble Alpes – Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble, Institut National de la Santé et de la Recherche Médicale - INSERM : U1038, Université Grenoble Alpes – France

[6] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France (MaIAGE) – Institut National de la Recherche Agronomique - INRA – Domaine de Vilvert, 78352 Jouy-en-Josas, France

[7] Exploring the Dynamics of Proteomes (EDyP), BGE/U1038, INSERM/CEA/Université Grenoble Alpes (CEA/DRF/BIG/BGE/EDYP) – Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble – 17 rue des Martyrs, 38054, Grenoble, France, France

Background: With the increased simplicity associated with producing MS-based proteomics data, the bottleneck has now shifted to the annotation and the exploration of large lists of proteins for biological interpretation purpose. As a joint effort between the French proteomics infrastructure (ProFI) and the French bioinformatics Institute (IFB), the ProteoRE (Proteomics Research Environment) primary aims to meet this need by centrally providing an online research service enabling biologists/clinicians with few programming expertise to analyze their proteomics datasets. ProteoRE is built upon the Galaxy web-based technology, a scientific workflow system allowing for data integration, data and analysis persistence and providing interfaces for users to interact with tools dedicated to the functional and the visual analysis of proteomics datasets.

Methods: Using output files from quantitative proteomics software (e.g. export from MaxQuant, Proline...) as a starting point, tools and interfaces have been designed in close collaboration with biologists and clinicians on the basis of real case studies. ProteoRE's tools have been implemented either by reusing tools (from the Galaxy Tool Shed) or by wrapping Bioconductor packages and external code, then beta-tested by "na´ive" end-users.

*Speaker

Results: Two case studies have been considered: the first one consists in interpreting a proteins identification list from a human sample [Lacombe *et al.*, 2018] while the second entails selection of biomarkers candidates. Tools implementing the corresponding workflows include i. data manipulation (filtering, identifiers conversion, cross-comparison), ii. protein list annotation (using public data resources such as Human Protein Atlas, neXtProt, etc.), iii. functional and pathways analyses (GO terms frequencies, enrichment analysis, mapping to Reactome) along with graphical representations. A first version of ProteoRE integrating these tools and datasets with online support and tutorials is in free access: http://www.proteore.org

Conclusions: While Galaxy-based tools offer services for proteomics identification (e.g. MS data conversion, protein database tools, search algorithms), tools focusing on proteomics functional analysis are still lacking. ProteoRE is an attempt to fill this gap. In view of a better "multi-omics" integration for proteomics data with genomics, transcriptomics, and metabolomics that has become highly relevant for many applications in system biology and medicine, we encourage the Galaxy community to contribute to ProteoRE and hope that this will pave the way for the development of tools and/or workflows towards this direction.

Keywords: Proteomics data annotation, functional analysis, web server, computer application, Galaxy

iCONICS: a bioinformatics/biostatistics core facility dedicated to the analysis and integration of biomedical multimodal data

Ivan Moszer *† [1], Aurélien Béliard *

[1], Thomas Gareau [1], Justine Guégan [1], Beáta György *

[1], Boris Labrador [1], François-Xavier Lejeune [1], Grégoire Limansky [1], Vincent Perlbarg [1], Arthur Tenenhaus [1,2], Damien Ulveling [1], Anissa Zareb [1], Stanley Durrleman [1,3]

[1] Bioinformatics/Biostatistics Core Facility (iCONICS) – Institut du Cerveau et de la Moelle épinière (ICM), CNRS UMR7225, INSERM U1127, UPMC – Hôpital Pitié-Salpêtrière, 47 boulevard de l'Hôpital, CS 21414, 75646 PARIS Cedex 13, France

[2] Laboratoire des Signaux et Systèmes (L2S) – CNRS, CentraleSupélec, UPMC, Univ Paris Sud – CentraleSupélec, 3 rue Joliot Curie, 91190 Gif sur Yvette, France

[3] ARAMIS Lab – Institut National de Recherche en Informatique et en Automatique – Hôpital Pitié-Salpêtrière, 47 boulevard de l'Hôpital, CS 21414, 75646 PARIS Cedex 13, France

The iCONICS core facility is member of a set of technological platforms gathered at the Institut du Cerveau et de la Moelle épinière (ICM), a private state-approved non-profit foundation dedicated to basic and clinical neuroscience research. Part of the largest European hospital (Pitié-Salpêtrière – with a major neurological department), the ICM is designed to understand how the brain and the spinal cord gives rise to mental life, behavior, and movements both under normal conditions and in diseases such as: Alzheimer, Parkinson, multiple sclerosis, epilepsy, depression or paraplegias. iCONICS is supported by the IHU-A-ICM program (Paris Institute of Translational Neurosciences).
The iCONICS core facility provides scientists and clinicians with a technical and analytical support for bioinformatics and biostatistics components of their biomedical studies, and designs innovative methods and tools in that purpose.

These missions are implemented through four lines of activities:

· Support for data management (including curation, standardization, annotation and structuration strategies) is proposed, using dedicated community tools and databases (e.g., REDCap for clinical eCRF, XNAT and OMERO for neuroimaging and cell imaging data, resp.). An open data warehouse, tranSMART, is used to provide reporting and visualization features from the integration of translational research data (genetics/omics/imaging/phenotypes), enabling cross- and meta-analyses.

*Speaker

†Corresponding author: ivan.moszer@icm-institute.org

· Pipelines are built and deployed (using the *snakemake* workflow manager) to process a range of genetics and (epi)genomics data, from high-throughput sequencing analyses – gene panel, whole-exome sequencing (SNPs, CNVs, rare variants), RNA-seq (differential gene expression, non-coding RNA, single-cell), bisulfite-seq (methylation profile), ATAC-seq (chromatin accessibility), ChIP-seq (protein binding) – and array analyses – GWAS, transcriptomics, methylation.

· Basic support in statistical data analysis is provided, and advanced methods are designed to deal with the integrative analysis of multimodal and high-dimensional data (e.g., genetics/omics data, electrophysiology and neuroimaging data, clinical observations), namely a versatile framework called Regularized Generalized Canonical Correlation Analysis (RGCCA), and its sparse counterpart SGCCA, dedicated to the analysis of data sets structured in blocks of variables.

· Graphical tools are developed and deployed to help in the operation of methods and in the interpretation of complex data. For instance, Shiny/R applications are available to explore transcriptomics data or launch integrative analyses of multimodal data; Web-based services are proposed for filtering and querying gene variant data, or built for brain lesion characterization and outcome prediction of coma patients (e.g., following traumatic brain injury) using neuroimaging data analysis in clinical settings.

Relying on expert staff and specialized methods and tools, basic and advanced support in bioinformatics and biostatistics is thus provided ”on demand” (study design, data management/processing/integration/interpretation, and software development) for around 50 projects and requests of variable scale each year.

iCONICS is a component of the newly created Center of Neuroinformatics of the ICM, a virtual and distributed structure aiming at harmonizing and sharing best practices in data management and analytics across the Institute. With all aspects of neuroscience – from basic biology to clinical research – covered at the ICM, breaking down barriers between multidisciplinary domains is expected to promote innovative approaches using scientific computing and computational modeling.

iCONICS is also affiliated to local (Sorbonne Université réseau Omique) and national (Institut Français de Bioinformatique) networks of core facilities.

Keywords: neuroscience, clinical data, neuroimaging, data management, database and datawarehouse, high throughput sequencing, genetics and omics, biostatistics, multimodal data integration, graphical interfaces

pitfalls and tricks in data analysis

Samuel Granjeaud * [1]

[1] Centre de Recherche en Cancérologie de Marseille (CRCM) – Centre National de la Recherche Scientifique : UMR7258, Institut National de la Santé et de la Recherche Médicale : U1068, Institut Paoli-Calmettes : UMR7258, Aix Marseille Université : UM105 – 27 bd Leï Roure, BP 30005913273 Marseille Cedex 09, France

Data in biology are rich and become complex when trying to integrate them. Data analysis is based on knowledge, visualization and statistics, but filtering and transformation play important roles. Although many tools and recipes are available, using them wisely requires attention. In the following lines, I present some interesting questions and some opinioned approaches from my experience.
Color heatmap is a widely used representation. Because an image is understood by the brain directly and globally, the transformation applied to each feature (line) is very important for interpreting this result. Scaling is usually carried out by computing the overall variance of the feature which accounts for the noise variance and for the between group variance. Surprisingly, features with high difference between groups might not get the highest color contrast. Centering is usually carried out by computing the overall mean of the feature. This value might belong to different groups of samples from one feature to the next. Centering on a chosen group of samples synchronizes the reading of the heatmap across the features.

P-value is the probability of getting a higher score if the null hypothesis is true. It should be clear that it is not the probability that the null hypothesis is true. Even if you avoid the dirty dozen [Goodman], do you state the right null hypothesis? For example, do you choose the correct universe gene set for a gene enrichment calculation?

P-value could become highly statistically significant when working with big samples. But is the effect size of the tested phenomenon of any practical importance? How will it translate at the wet bench or the patient bed? Is the parameter under study the right one? For example, in a differential expression study, should you analyze the count of transcripts or the count of reads per transcript?

When considering p-value as a risk of false positive, it's easier to understand that the repetition of statistical tests leads to an increased overall risk. The False Discovery Rate is an approach to moderate it. Although there is no standard threshold such as the famous 5%, the FDR can be lowered by prior operations. Instead of testing all the features, features can be clustered and representatives selected. Testing only representatives does not change their p-value but reduces the number of tests, which usually reduces the false discoveries. Alternatively filtering features blindly with respect to the experimental design reduces also the number of tests. In sequencing approaches, filtering to remove low counts is now common [Chen et al.]. But filtering can be applied to remove features of low dynamic range [Burbon].

*Speaker

Preparing a meaningful representation, using statistics wisely and humbly, keeping in mind the biological or clinical effect size are the main points to collaborate with experimentalists in my everyday practice of bioinformatics.

References

A dirty dozen: twelve p-value misconceptions. S Goodman - 2008

From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. Y Chen, A Lun, GK Smyth - 2016
Independent filtering increases detection power for high-throughput experiments. R Bourgon, R Gentleman, W Huber - 2010

Keywords: heatmap, transformation, visualization, statistics, false discovery, filtering, hacking

Spark-ics, exploration de l'application de l'architecture Apache Spark à la bioinformatique

Axel Verdier * [1], Ludovic Legrand [1], Xavier Garnier [2], Erika Sallet [1], Alexandre Dehne-Garcia [3], Nicolas Lapalu [4], Martial Briand [5], Franck Dorkeld [3], Bernhard Gschloessl [3], Corinne Rancurel [6], Martine Da Rocha [6], Sébastien Carrère [1], Céline Noirot [7], Olivier Filangi [8,9], Jérôme Gouzy† [1]

[1] Laboratoire des Interactions Plantes-Microorganismes (LIPM) – Université dToulouse, Institut national de la recherche agronomique (INRA) : UMR441, CNRS : UMR2594 – 24 chemin de Borde Rouge - Auzeville CS 52627 31326 CASTANET TOLOSAN CEDEX, France

[2] DYnamics, Logics and Inference for biological Systems and Sequences (Dyliss) – IRISA, INRIA, Bretagne Atlantique – France

[3] Centre de Biologie pour la Gestion des Populations (CBGP) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement : UMR55, Centre international d´tudes supérieures en sciences agronomiques : UMR1062, Institut national de la recherche agronomique [Montpellier] : UMR1062, Université de Montpellier : UMR1062, Institut de Recherche pour le Développement : UMR1062, Institut national d'études supérieures agronomiques de Montpellier, Centre international d´tudes supérieures en sciences agronomiques : UMR1062, Centre international d´tudes supérieures en sciences agronomiques : UMR1062, Centre international d´tudes supérieures en sciences agronomiques : UMR1062 – 755 avenue du Campus Agropolis, 34988 Montferrier sur Lez, France

[4] BIOlogie GEstion des Risques en agriculture (BIOGER) – Institut National de la Recherche Agronomique : UMR1290, AgroParisTech – Campus AgroParisTech BP01 F-78850 Thiverval-Grignon, France

[5] Institut de Recherche en Horticulture et Semences (IRHS) – Université d'Angers, Institut National de la Recherche Agronomique : UMR1345, Agrocampus Ouest – AGROCAMPUS OUEST, UMR1345 IRHS, F-49045 Angers, France, France

[6] Institut Sophia Agrobiotech (ISA) – Institut national de la recherche agronomique (INRA) : UMR1355, CNRS : UMR7254, Université Côte d'Azur (UCA) – 06903 Sophia Antipolis Cedex, France

[7] Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT INRA) – Institut National de la Recherche Agronomique : UR875 – Chemin de Borde Rouge, 31320 Castanet Tolosan, France

[8] Institut de Génétique, Environnement et Protection des Plantes (IGEPP) – Institut National de la Recherche Agronomique : UMR1349, Universite de Rennes 1 : UMR1349, Agrocampus Ouest : UMR1349 – Domaine de la Motte au Vicomte BP 3532735653 Le Rheu, France

[9] Plateforme bio-informatique Genouest (Genouest) – genouest – IRISA/INRIA Plateforme GenOuest Campus de Beaulieu 35042 Rennes cedex, France, France

Les architectures bioinformatiques actuelles sont majoritairement basées sur l'exploitation de clusters de calculs, solution versatile, permettant l'utilisation de nombreux outils, de moyens de calcul conséquents et l'accès à de nombreuses bases de données externes. A ce jour, les outils disponibles sur le cloud répondent à des besoins d'analyses ponctuels. L'exploitation des archi-

*Speaker

†Corresponding author: Jerome.Gouzy@inra.fr

tectures issues du " big data " comme les clusters " Apache Spark " [1] reste balbutiante en bioinformatique. Contrairement à la génération précédente basée sur Hadoop qui, comme les clusters de calcul, nécessite des accès disques à chaque étape d'un pipeline, Spark implémente les "Resilient Distributed Datasets" qui permettent d'enchaîner en mémoire les différentes étapes. C'est potentiellement une source d'optimisation critique du problème le plus fréquent dans les pipelines bioinformatiques intégrant successivement de nombreux outils et filtres. Depuis 2016, un financement de plusieurs entités INRA a permis la mise en place d'un cluster Spark dédié a l'exploration de cette technologie dans le cadre du groupe de travail "spark-ics"

Dans un premier temps nous présenterons l'architecture du cluster Spark de 8 nœuds et 416 cœurs que nous avons mis en place. Nous présenterons les différentes couches logicielles de l'architecture Spark (scheduler YARN, API, etc.) ainsi que l'écosystème de la distribution MapR [2]. MapR implémente un système de fichiers distribué permettant un accès POSIX et via l'API HDFS.

En outre, depuis début 2017, un nombre croissant de publications évaluent l'apport de Spark sur des problèmes de bioinformatiques classiques (BLAST, alignements multiples, métagénomique, etc) avec un niveau de maturité de plus en plus pertinent. Ainsi, Spark-hit [3] permet d'instancier un cluster Spark sur le cloud Amazon pour répondre à des problématiques bioinformatiques (métagénomique, détection de variants) et GATK [4] (détection de variants) a été réécrit par les équipes du Broad Institute pour exploiter les clusters Spark. Nous avons évalué ces deux derniers outils et nous présenterons d'une part leurs caractéristiques techniques les plus intéressantes ainsi qu'un comparatif de performance entre une exécution du pipeline GATK sur notre cluster de calcul en mode SGE et en mode Spark.

Au delà de la scalabilité, l'intérêt de migrer sur une architecture " big data " des chaines de traitements informatiques est également de disposer au niveau programmatique des outils et méthodes d'apprentissage automatique (Machine Learning) et ainsi de pouvoir les intégrer dès que nécessaire dans le processus sans avoir à passer par l'exécution d'outils externes et d'avoir éventuellement à gérer le parallélisme (ou de tout recoder!). Ainsi, la librairie ML/MLlib implémente nativement sur Spark de très nombreuses méthodes de Machine Learning. Nous présenterons un pipeline Spark de détection de contaminants dans les assemblages génomiques. Ce programme scala couple des étapes classiques de processing de fichiers et l'implémentation des 'Random Forests' de la librairie ML pour prendre les décisions.

Matei Z. et al. 2011. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. http://www.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-82.html

MapR: https://mapr.com/

Liren Huang, Jan Kŕuger, Alexander Sczyrba. 2017. Analyzing large scale genomic data on the cloud with Sparkhit. Bioinformatics.

McKenna et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research.

Keywords: spark, cluster, parallélisme, machine learning, big data

Groupe de travail sur les Métiers de la Bioinformatique (MetBIF)

Hélène Chiapello * [1], Samuel Mondy [2], Morgane Thomas-Chollier [3]

[1] Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE) – Institut national de la recherche agronomique (INRA) : UR1404 – France
[2] UMR Agroécologie – Institut national de la recherche agronomique (INRA) – Dijon, France
[3] Institut de Biologie de l'Ecole Normale Superieure (IBENS) – INSERM 1024 CNRS 8197 – 46 rue d'Ulm 75005 Paris, France

Les métiers de la bioinformatique s'exercent le plus souvent dans des contextes pluri ou inter-disciplinaires en rapide évolution. Il n'existe pas à ce jour de panorama clair et partagé des différents métiers de la bioinformatique, c'est un manque important pour notre communauté dans sa globalité (étudiants, enseignants, secteur professionnel public et privé).
La SFBI (Société Française de Bioinformatique) a lancé début 2018 un groupe de travail sur les métiers de la bioinformatique (MetBIF) qui a pour objectif de dresser ce panorama des différents profils de métiers de la bioinformatique (ingénieurs, chercheurs, enseignants-chercheurs) et des spécificités liées aux différents contextes dans lequel s'exercent leurs activités (public ou privé, domaines d'application, recherche ou service,...).

Une enquête a été lancée auprès de la communauté entre le 10 avril et le 16 mai 2018 afin de mieux caractériser les activités et compétences des différents types de métiers de la bioinformatique, en tenant compte du contexte dans lesquels ils s'exercent.

Une première journée de travail et de réflexion est organisée le 31 mai 2018 juste avant la 2ème édition de REBIF (Réseau des enseignants en Bioinformatique), regroupant une vingtaine de personnes représentant différents métiers et contextes professionnels de la bioinformatique.

Ces deux actions vont permettre de commencer à construire une typologie des différents métiers de notre communauté. Ces profils seront mis en regard des différentes formations actuellement proposées par les universités, écoles et organismes de formation.

Ce panorama des métiers permettra de mieux présenter les débouchés possibles dans le domaine de la Bioinformatique et bénéficiera à l'ensemble de notre communauté.
Le poster présenté à JOBIM détaillera les résultats de l'enquête et le bilan de la journée de travail.

Keywords: métiers, SFBI, MetBIF, formation, recherche, ingénierie, service

*Speaker

Straightforward finding of differential expressions through intensive randomization in transcriptomic studies

Dorota Desaulle * [1], Céline Hoffmann [2], Pascal Bigey [2], Bernard Hainque [2], Yves Rozenholc† [3]

[1] Université Paris Descartes - Paris 5 - EA4064 (UPD5) – EA4064 – 12, rue de l´cole de Médecine - 75270 Paris cedex 06, France
[2] Université Paris Descartes - Paris 5 - PSL (UPD5) – UMR CNRS 8258, INSERM U1022, PSL Research University, Chimie ParisTech – CNRS, Institut de Recherche de Chimie Paris, 75005, Paris, France – 12, rue de l´cole de Médecine - 75270 Paris cedex 06, France
[3] Université Paris Descartes - Paris 5 (UPD5) – Institut de Recherche pour le Développement - IRD (FRANCE) : UMRIRD 216 – 12, rue de l´cole de Médecine - 75270 Paris cedex 06, France

Transcriptomic data measures proportions of transcripts relative to all amount of RNA in a biological sample. One major issue in transcriptomic studies is that, even if the amount of analyzed tissue is set precisely by the biologist, only a fraction of this quantity reacts during the analysis. As a consequence, an absolute quantification is not available directly from the observations.
Much effort has been done to control this intrinsic variability by adjusting each sample by a multiplicative factor before any differential comparisons between samples (e.g. case-control study). This adjustment is called normalization. Apart from inadequate methods for differential studies based on library size or housekeeping genes, normalizations try to find a proper subset of invariants to estimate the scaling factor without any prior knowledge. Although, such strategies may be justified in some contexts, they can be shown to fail on simple counter-examples by adjusting expression variabilities across the conditions and/or the different expressions.

Under the assumption that the majority of analyzed expressions is invariant, we propose a new procedure for finding differential expressions. This procedure is straightforward in the sense that it is not preceded by a previous "good" normalization step. Instead, the findings are obtained consecutively to a sequence of normalizations, each obtained by selecting at random a small set of expressions, regardless their quality.

Keywords: normalization, differential expressions, differential analysis, transcriptom, RNA, miRNA

*Speaker

†Corresponding author: yves.rozenholc@parisdescartes.fr

A tool to interactively query, mine, and visualise Ribosome profiling data.

Damien Paulet *† [1,2], Alexandre David [3], Eric Rivals *

[2,4]

[1] Institut de Biologie Computationnelle (IBC) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut National de la Recherche Agronomique, Institut National de Recherche en Informatique et en Automatique, Université de Montpellier, Centre National de la Recherche Scientifique – 860 rue de St Priest, 34095 Montpellier, France

[2] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) – CNRS : UMR5506, Université Montpellier II - Sciences et Techniques du Languedoc – Montpellier, France

[3] Institut de Génomique Fonctionnelle (IGF) – CNRS, Institut National de la Santé et de la Recherche Médicale - INSERM – Team "Signaling and Cancer" / Oncology Department 141 rue de la Cardonille 34094 Montpellier cedex 5 FRANCE, France

[4] Institut de Biologie Computationnelle (IBC) – Univrsité de Montpellier – Montpellier, France

The last main phase of gene expression is the translation of messenger RNAs by the ribosomal machinery. For decades translation was viewed as a process that systematically produces one protein for any mRNA that was engaged in translation. Consequently, it was thought to have little impact on the control of gene expression. Recent works have demonstrated its role in the selection of translated mRNAs and as a key factor of protein level. For example, alternative, instead of classical, Open Reading Frames (ORFs) can be translated in certain conditions, even ORFs located in regions upstream from the translation start sites have been discovered. These achievements were made possible using Ribosome profiling (a.k.a. Ribo-seq), a sequencing assay that captures the fragments of RNAs protected by the ribosome, which are called Ribosome Protected Fragments (RPFs). One Ribo-seq experiment typically delivers tens of millions of reads extracted from the whole transcriptome.

Although some tools have been developed for the initial processing of Ribo-seq data, we lack programs enabling an interactive exploration and mining of such data. We need computational tools working that helps the user to mine potentially new or alternative ORFs in Ribo-seq datasets. Potential applications include functional annotation of transcriptomes but also discovery of potential novel peptides.

We present a tool allowing dynamic interactive visualisation of Ribo-seq profile along a RNA sequence. More importantly, with it the user has the capacity to perform queries to select RNAs/genes of interest. Three main kinds of queries are implemented:

- query RNAs on profile coverage over specific transcript regions. For example, select all genes showing a strong Ribo-seq profile in an untranslated region (e.g., in 5'UTR).
- detect RNAs/genes with the strongest changes in Ribo-seq profile across two conditions.
- identify RNAs/genes exhibiting a change of Ribo-seq profile in user defined regions. Both the regions and quantitative changes are set by the user as criterion of selection. Those criteria can

*Speaker

†Corresponding author: damien.paulet@lirmm.fr

be saved with the session for later use and recovered in another session.

In practice, the user loads the reference transcriptome, the count data resulting from the mapping phase. He/she can then perform specific checks to verify the data quality and select the appropriate fragment lengths or position offset according to the expected trinucleotidic periodicity. After the verification phase, the user can proceed with queries and selection to mine the Ribo-seq data and find genes, potential ORFs, or variations of interest. After the verification phase, the user can proceed with queries and selection to mine the Ribo-seq data and find genes, potential ORFs, or variations of interest. The user can save the coverage plots or various diagrams as images in a vectorial format, as well as the data of his session for reuse.

In conclusion, our tool provides an interactive way to mine and visualise Ribo-seq profiles on any reference set of transcripts.

Demande: communication orale et démonstration

Keywords: ribo, seq, NGS, traR, uORF, profile, RPKM, read, coverage, visualisation

IRSOM, a reliable identifier of ncRNAs based on supervised Self-Organized Maps with rejection

Ludovic Platon[1,2], Farida Zehraoui[1], Abdelhafid Bendhamane[2], and Fariza Tahi[1]

[1] IBISC, Univ. Evry, Université Paris-Saclay, Evry, 91025, France.
[2] Institute of Plant Sciences Paris-Saclay, INRA, CNRS, Université Paris-Sud, Université d'Evry, Université Paris-Diderot, Orsay, France.

1 Introduction

Non-coding RNAs (ncRNAs) are transcripts that do not encode for proteins, contrary to coding RNAs. They are of different classes (ribosomal RNAs, transfer RNAs and microRNAs for example) and play important roles in many biological processes and are involved in many diseases such as cancer [1].

There are multiple tools to discriminate coding and non-coding RNAs [12, 13, 11, 17, 4, 7, 15, 16, 14, 9]. A basic idea to separate coding and non-coding transcripts is to evaluate the coding potential of a transcript with its Open Reading Frame (ORF) or its sequence composition [6].

The most popular tool, named CPC [11], is based on this idea. The authors built a model based on an SVM algorithm using several sequence features like ORFs quality, and on BLASTX results against a protein database. But this approach has an important drawback of time consuming. Recently, a new version of CPC, that is alignment-free, has been proposed by the same group. The new method, called CPC2 [9], uses SVM technique to build a model from four features: the Fickett score [5], the length, integrity and isoelectric point of the longest ORF.

CPAT [17], CNCI [15] and PLEK [13] are three other existing methods that are alignment-free and show comparable or better result than CPC in their respective articles. These methods use features extracted from the sequences such as the Fickett score, the maximal ORF length and coverage, sequences motifs (adjoining nucleotide Triplets (ANT), Hexamer frequencies or k-mer for example).

Here we present IRSOM, a new alignment-free method for discriminating non-coding and coding RNAs. IRSOM is a supervised classifier composed of a Self-Organizing Map (SOM) and a perceptron layer which is fully connected to the SOM. IRSOM uses several features that are related to the sequence statistics (k-mers motifs frequencies, codon position biases, nucleotide frequencies and GC content) and the putative ORFs (coverage of the longest ORF, ORFs coverage distribution, start and end codon distribution, ORF frequency, ORF length and the frame bias). We also associated a rejection option to IRSOM. The rejection allows to keep reliable predictions and to abstain in the situations where the predictions are unreliable. Moreover, by combining the rejection option with the SOM, we are able to visualize and analyse the rejected transcripts. For example, analysing the ORF features profiles in the SOM allows to highlight the known differences between the coding and the non-coding RNAs and also shows the ambiguous characteristics of the rejected transcripts.

2 Method

Self-Organizing Map (SOM) [10] is a neural network that is able to cluster and visualize high dimensional data. By using an unsupervised competitive learning algorithm, SOM is able to produce a map representing the input space. IRSOM is a three layers neural network composed of an input layer that represents the input data, a hidden layer which corresponds to a SOM [10], and an output layer (supervised layer), that consists of two perceptrons which are fully connected to the neurons of the SOM with forward connections. Moreover, we extend the perceptron layer with a rejection option where the ambiguous predictions are rejected.

The neural network is trained using a forward and backward propagation. The activation of the neurons in the hidden layer are propagated to the output layer. And the error is back propagated in the neural network in order to optimize its weights.

Let a set of input data $X = \{x_1, x_2, ..., x_n\}$ and their corresponding labels $Y = \{y_1, y_2, ..., y_n\}$ such that

$y_i \in \{0,1\}^2$ is a vector representing the label of x_i. An element $y_i \in Y$ is defined such that:

$$y_i = \begin{cases} [1,0] & \text{for coding RNAs} \\ [0,1] & \text{for non-coding RNAs} \end{cases} \tag{1}$$

The activation a_{iu} of the unit u in the hidden layer depends on u and its neighbors u' such that:

$$a_{iu} = \sum_{u' \in U} exp\left(-\frac{1}{2} \parallel x_i - w_{u'} \parallel^2\right) \sigma_t(u', u)$$

$$\sigma_t(u', u) = exp\left(-\frac{d(u', u)^2}{\alpha \times \left(1 - \frac{t}{T}\right) \times r}\right)$$

where $d(u', u)$ is the Manhattan distance between the neuron u' and u and α is a constant. The output o_l $\forall l \in \{0,1\}$ of the two perceptrons are computed as:

$$o_{il} = sig(act_{il}) \tag{2}$$

where $act_{il} = \sum_u w^{out}_{ul} a_{iu} + b_l$ and sig is the sigmoid function. w^{out}_{ul} is the connection between the perceptron l and the map unit u and b_l is the bias of the perceptron l.

With the output o_l we compute the loss function $L()$, which consists of the cross-entropy cost function $C()$ and a L2-norm regularization term such that:

$$L(Y, O) = C(Y, O) + \lambda \sum_u \parallel w^{out}_u \parallel^2 \tag{3}$$

where O is a vector containing the ouput of the perceptrons, λ is the parameter which controls the importance of the regularization term, and

$$C(Y, O) = -\frac{1}{N} \sum_i \sum_l y_{il} \ln(o_{il}) \tag{4}$$

By using the loss function $L()$ we determine the gradient of the weights in the neural network. Each weights are update using the momentum optimizer such that: The weights of our neural network are optimized using the momentum optimizer such that:

$$w(t+1) = w(t) - \mu_1 \times \left(\mu_2 \times acc_w + \frac{\partial L(Y, O)}{\partial w}\right) \tag{5}$$

where w is a weight of the neural network, acc_w represents the sum of the gradient for this weight over the iterations, and μ_1 and μ_2 are constants controlling respectively the learning rate and the importance of the accumulation.

To improve the reliability of our method and identify the ambiguous transcripts, we use one of the rejection approaches proposed in [8]. The greater the difference between o_{i0} and o_{i1} is, the greater is the confidence in the prediction. By following the second rejection method in the article [8], we can improve the reliability of the prediction by rejecting the ambiguous classifications. We are able to define a classifier with rejection option called $\psi(x_i)$ such that:

$$\psi(x_i) = \begin{cases} -1 & \text{if } |o_{i0} - o_{i1}| < \beta \\ \arg\max_l \ o_{il} & \text{otherwise} \end{cases} \tag{6}$$

where β is the rejection threshold. When the absolute difference value between o_{i0} and o_{i1} is lower than a threshold β, the prediction is rejected and set to -1.

The parameter β is application dependent. For certain applications we may want a high β in order to have the most reliable predictions but in an exploratory analysis, we may use a smaller β in order to keep more predictions even if they are potentially misclassified.

3 Results

We evaluated our method with coding and non-coding RNAs coming from several species. In order to cover a large spectrum of species from different reigns, we selected RNAs from Human, Mouse, Oryza sativa, Arabidopsis thaliana, Zebrafish, Escherichia coli, Saccharomyces cerevisae and Drosophila. The sequences are extracted from Ensembl [18], GENCODE [3] and RNAcentral [2].

We show the performance of our tool by comparing it to four classical ncRNA identification tools which are CPAT [17], CPC2 [9], CNCI [15] and PLEK [13]. We performed two types of performance analysis: a cross-validation analysis and a prediction analysis. For this purpose, each of the different datasets, has been divided into two subsets (of same size) where the first subset is the training set and the second is the prediction set. In the cross-validation, we evaluate our tool IRSOM in order to measure the impact of the rejection threshold on the different datasets. We therefore performed a 10-fold cross-validation with IRSOM on the training set, and then a prediction with the different tools, CPAT, CPC2, CNCI, PLEK and IRSOM on the prediction set. IRSOM and CPAT are trained on each of the considered species, as well as on all species together (cross-species model). In this last case we note the two tools as IRSOM_cross and CPAT_cross respectively. We measure the classification performance using three measures:

- Accuracy: represents the percentage of correctly classified RNAs, it is defined as follows:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

- Sensitivity: measures the rate of true positives:

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

- Specificity: measures the rate of true negatives:

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

where TP are the true positives, TN are the true negatives, FP are the false positives and FN are the false negatives. Here the positive class represents the non-coding RNAs and the negative one the coding RNAs. In the case of IRSOM, the TP, TN, FP and FN are computed on the non rejected data.

The cross-validation results (Figure 1) show good performance of our method on all the datasets. We can also see that the performance of IRSOM increases when we increase the rejection threshold. For the Drosophila, Human, Mouse and Oryza sativa, the improvement is stronger than on the other species due to the higher amount of rejected value. These differences can be explained by the largest presence of ambiguous transcripts in these datasets.

In order to define the rejection threshold of IRSOM, we use the cross-validation results. For the cross-species model of IRSOM, we set the rejection threshold at 0.7. This threshold gives good performance for all species (Accuracy greater than 0.975) with a reasonable rejection rate (less than 20% for all species). For the species specific model, we set a threshold for each species. We set the thresholds to values that show the highest performances. For the S. cerevisiae and E. coli, we set a threshold of 0.1 and 0.2 respectively. For the plants, we set a threshold of 0.6 for the Arabidopsis and 0.8 for the Oryza sativa. The eucaryotes species have a wider range of threshold. We set a threshold of 0.5 for the Zebrafish, 0.75 for the Drosophila and Mouse and 0.8 for the Human. For all the species, we have at most 30% of the predictions that are rejected. For most of them, we have a reject rate lower than 20% (A. thaliana thaliana and Oryza sativa) or even 10% (E. coli, Zebrafish and S. cerevisiae).

Figure 2 shows the prediction performance of IRSOM and the benchmarked tools. In the case of IRSOM, we compute a model on the whole training set and use the rejection threshold defined with the cross-validation results. The obtained results show a good performance of our tool IRSOM compared to the other tools. IRSOM exceeds 0.95 in accuracy for all the species. Compared to CPAT, the only tool we succeeded to retrain on our data and the second best tool, IRSOM shows slightly better results for all considered species. Furthermore, the two models of CPAT show the same performance on all datasets as for the two models of IRSOM (except on the Human). Finally, IRSOM gives comparable execution time compared to the faster tools (CPC2 and CPAT). For example, on the Human dataset, CPAT, CPC2 and IRSOM achieved their prediction in 60, 57 and 82 seconds respectively when CNCI and PLEK take 2 747 and 370 seconds respectively.

In order to understand why some transcripts are rejected, we investigate the prediction by visualizing the SOM prototypes and the distribution of the labels in the SOM for the Human dataset. With the projection properties of SOM, we can extract the profiles of the transcripts that are rejected by taking the the weight vector of their closest neurons in the hidden layer.

The predicted label repartitions in the SOM show that the rejected transcripts are in the overlapping area between the coding and non-coding transcripts. The ORF profiles of the rejected transcript show an ORF coverage of 0.5 which mean that the longest ORF find cover half of the transcript length. Moreover, they show a high ORF frequency like the non-coding transcripts. The average ORF coverage with the high ORF frequency suggests that these transcripts have coding sequences that are not stable as the other coding transcripts. This suggests that these transcripts are potentially degraded coding transcripts.

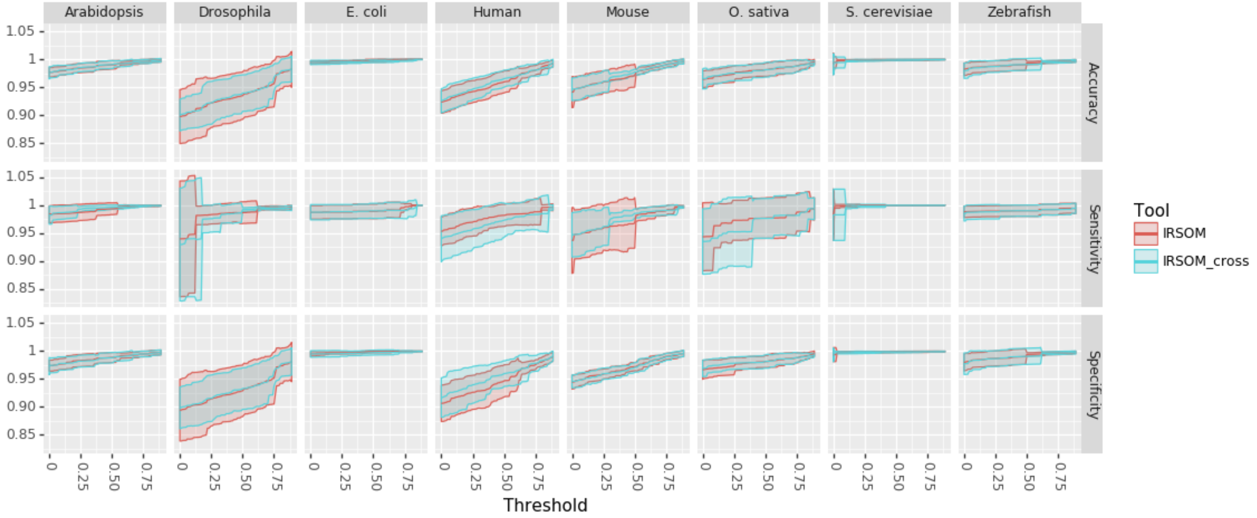


Figure 1: IRSOM cross-validation performance (mean ± standard deviation) in regard to the rejection threshold for all the datasets.

4 Conclusion

Our tool, called IRSOM, is able to accurately discriminate coding and non-coding RNAs. Furthermore, with our rejection option, we are able to identify the ambiguous transcripts and analyse them with the SOM. Compared to the state of art, our tool gives the best results on several species of different reigns. It gives also good time computing for small and large datasets.

One of our future work is to extend our algorithm in order to take into account different heterogeneous data sources. The sources could be numerical vectors or more complex data like graphs. By doing so, we will be able to use new features such as secondary structures or epigenetic profiles for the classification task. These new features will be used to classify ncRNAs into different classes corresponding to ncRNA types, like for example transfert RNA (tRNA), ribosomal RNA (rRNA), microRNA (miRNA) or piwi RNA (piRNA).

References

[1] Nenad Bartonicek, Jesper L V Maag, et al. Long noncoding RNAs in cancer: mechanisms of action and technological advancements. *Molecular Cancer*, 15(43), 2016.

[2] The RNAcentral Consortium. RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Research*, 45(D1):D128–D134, jan 2017.

[3] Thomas Derrien, Rory Johnson, et al. The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. *Genome research*, 22(9):1775–1789, 2012.

[4] X N Fan and S W Zhang. lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning. *Molecular bioSystems*, 11(3):892–897, 2015.

[5] James W Fickett and Chang-Shung Tung. Assessment of protein coding measures. *Nucleic acids research*, 20(24):6441–6450, 1992.

[6] Gali Housman and Igor Ulitsky. Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1859(1):31–40, jan 2016.

[7] Long Hu, Zhiyu Xu, et al. COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic acids research*, 45(1):e2, jan 2017.

[8] Hisao Ishibuchi and Manabu Nii. Neural networks for soft decision making. *Fuzzy Sets and Systems*, 115(1):121–140, 2000.

[9] Yu-Jian Kang, De-Chang Yang, et al. Cpc2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic acids research*, 45(W1):W12–W16, 2017.

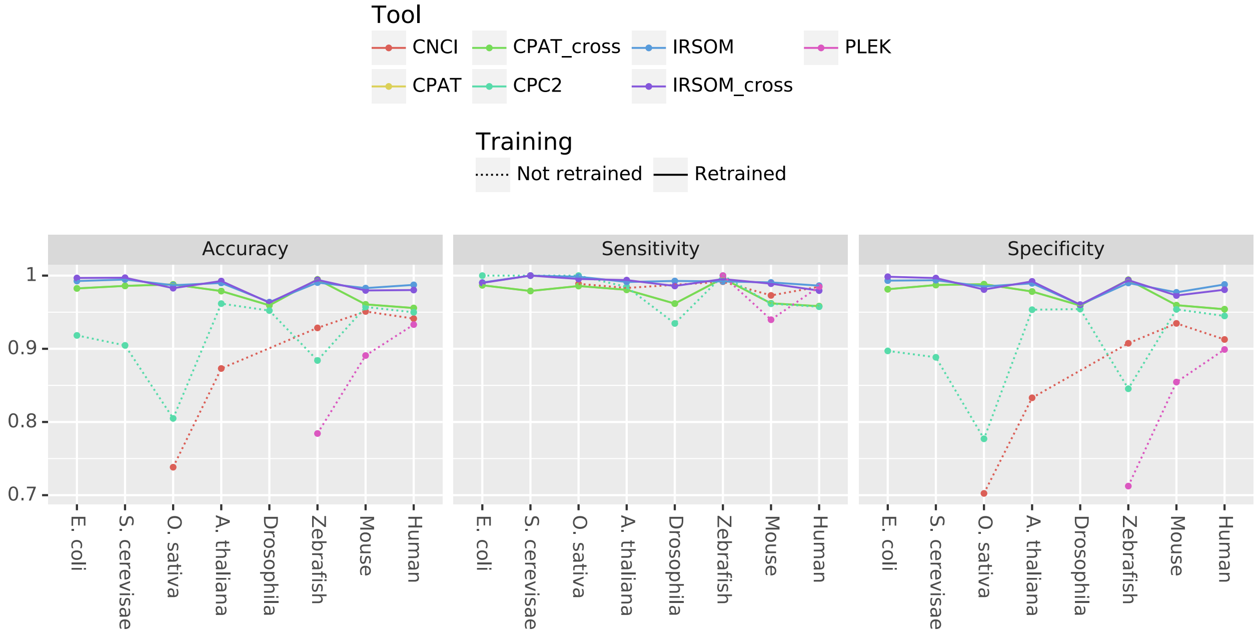


Figure 2: Accuracy results obtained by CNC2, CNCI, CPAT, PLEK and IRSOM on each of Human, Mouse, A. thaliana thaliana, Oryza saliva, Zebrafish, Escherichia coli, Saccharomyces cerevisiae and Drosophila species. CPAT_cross and IRSOM_cross designate respectively CPAT and IRSOM when used with the cross-species model. CPAT and IRSOM are by default used on each species with the corresponding model. Two models for CNCI are available, one for vertebrate and one for plants. CPC2 was trained on Human protein and ncRNA in GENCODE. And PLEK was trained on Human data.

[10] Teuvo Kohonen. *Self-Organizing Maps - third edition.* Springer-Verlag Berlin Heidelberg, 2001.

[11] Lei Kong, Yong Zhang, et al. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, 35(SUPPL.2):345–349, 2007.

[12] Supatcha Lertampaiporn, Chinae Thammarongtham, et al. Identification of non-coding RNAs with a new composite feature in the Hybrid Random Forest Ensemble algorithm. *Nucleic Acids Research*, 42(11):1–12, 2014.

[13] Aimin Li, Junying Zhang, et al. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics*, 15(1):311, 2014.

[14] M. F. Lin, I. Jungreis, et al. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27(13):i275–i282, jul 2011.

[15] Liang Sun, Haitao Luo, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research*, 41(17), 2013.

[16] Giovanna M M Ventola, Teresa M R Noviello, et al. Identification of long non-coding transcripts with feature selection: a comparative study. *BMC bioinformatics*, 18(1):187, mar 2017.

[17] Chunyu Wang, Leyi Wei, et al. Computational Approaches in Detecting Non-Coding RNA. *Current Genomics*, 14:371–377, 2013.

[18] Daniel R Zerbino, Premanand Achuthan, et al. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2018.

Author Index

ORGANIZED BY

sfbi
société française de bioinformatique

Inserm
La science pour la santé
From science to health

cnrs

Aix*Marseille
université
Initiative d'excellence

ifb
INSTITUT FRANÇAIS
DE BIOINFORMATIQUE

GdR BiM

PARTNERS & SPONSORS

A*Midex
Initiative d'excellence Aix-Marseille

aviesan
alliance nationale
pour les sciences de la vie et de la santé

cancéropôle
Provence-Alpes-Côte d'Azur
le propulseur régional des recherches
et innovations anticancers

EMBL-EBI

FONDATION ARC
POUR LA RECHERCHE
SUR LE CANCER
Reconnue d'utilité publique

FRANCE GÉNOMIQUE

GDR
GÉNOMIQUE
ENVIRONNEMENTALE

GenoScreen
Innovative Genomics

GRAND LUMINY
Technopôle

INRA
SCIENCE & IMPACT

INSTITUT PAOLI-CALMETTES
Centre régional de lutte contre le cancer Provence-Alpes-Côte d'Azur

OMIC
TOOLS

RÉGION
SUD
PROVENCE
ALPES
CÔTE D'AZUR

DÉPARTEMENT
BOUCHES
DU RHÔNE

VILLE DE
MARSEILLE
www.marseille.fr