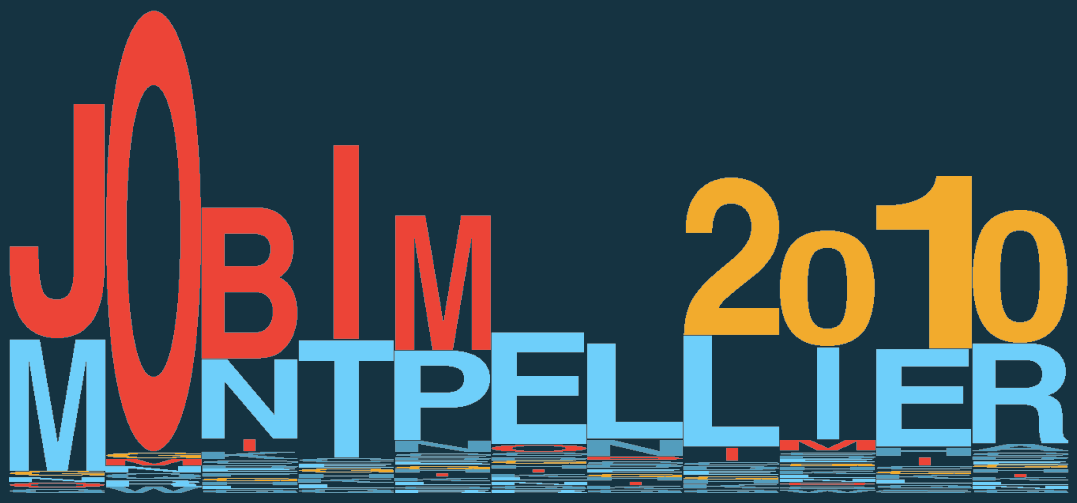
Evaluating Genome Browsers Using a Software Qualification Method Evaluation

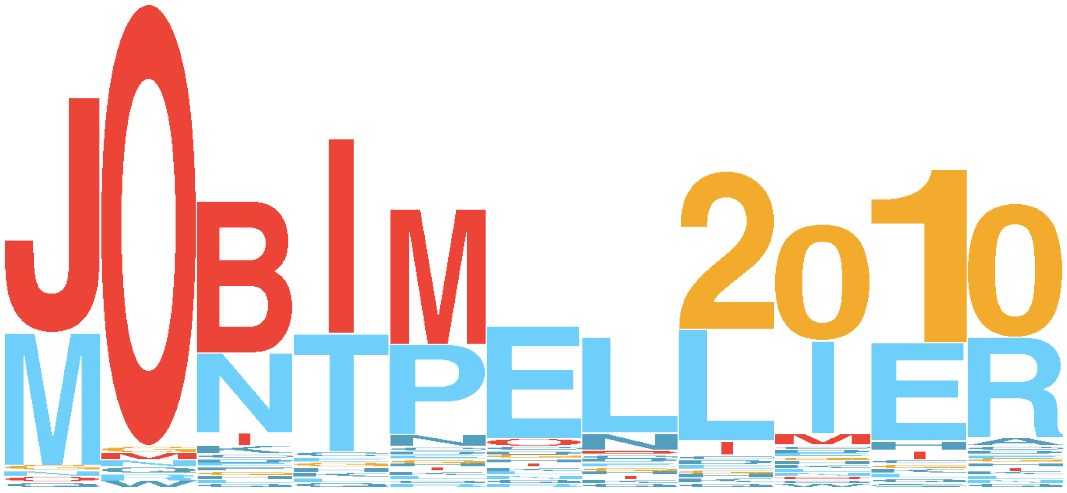
Thomas Lacroix, Valentin Loux, Annie Gendrault-Jacquemard Gendrault, Jean-François Gibrat, Helene H. Chiapello



Journées Ouvertes Biologie Informatique Mathématiques

Montpellier, 7 - 9 septembre 2010

Éditeurs
Olivier Gascuel
Marie-France Sagot



Journées Ouvertes Biologie Informatique Mathématiques

Montpellier, 7 - 9 septembre 2010

Éditeurs
Olivier Gascuel
Marie-France Sagot

Journées Ouvertes de Biologie, Informatique et Mathématiques

IV+xiv+176 pages.

Le site web de JOBIM2010 est accessible à l'adresse :
http://www.jobim2010.fr/

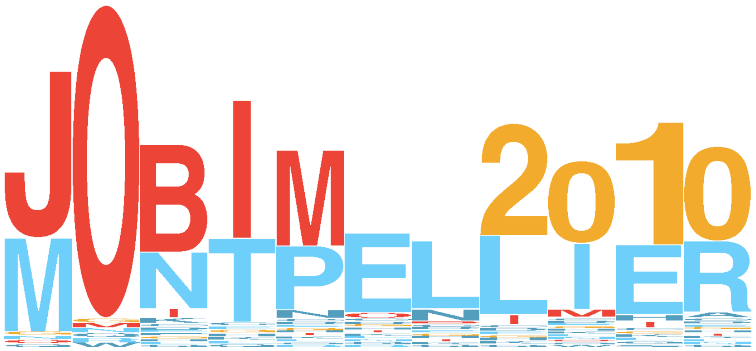
Les résumés des posters présentés lors de la conférence sont disponibles à l'adresse :
http://www.jobim2010.fr/?q=fr/node/55

Les vidéos des conférences invitées de la conférence sont disponibles à l'adresse :
http://www.jobim2010.fr/?q=fr/node/49

Impression: 19 juillet 2010

Réalisation et mise en pages par Alban MANCHERON.
Graphisme de la couverture de Laurent BRÉHÉLIN.

Ce recueil d'actes a été imprimé par l'imprimerie Bonniol.
http://www.imprimeriebonniol.com/
Zone artisanale Parc 2000,
126, rue Claude François
34080, Montpellier.



La conférence JOBIM est née il y a 10 ans à Montpellier, où elle revient cette année. C'est un lieu de rencontre ouvert à toutes les personnes travaillant aux frontières de la biologie, de l'informatique, des mathématiques et de la physique, afin de favoriser les échanges scientifiques et d'encourager l'expression des jeunes chercheurs. Les grands thèmes sont liés à la génomique, la bioinformatique structurale, la biologie des systèmes et l'analyse des données d'expression, l'évolution et la phylogénie, les bases de données et de connaissances, l'algorithmique et la modélisation, en particulier issue des probabilités et des statistiques. Mais la discipline se renouvelle et voit de nouveaux champs s'ouvrir, par exemple en analyse d'images, en génétique des populations ou du côté de l'écoinformatique. Elle bénéficie de données toujours plus abondantes et diverses, notamment de séquences grâce à l'amélioration spectaculaire des techniques de séquençage. Ces données à grande échelle permettent de répondre à de nouvelles questions, liées à l'épigénétique par exemple, mais elles imposent aussi de revoir les méthodes et les techniques.

Nous avons reçu cette année 66 soumissions, 16 ont été retenues pour des présentations longues et 26 pour des présentations courtes, auxquelles s'ajoutent les conférences invitées de Raphaël Guérois, Jean-Christophe Olivo-Marin, Luis Quintana-Murci, Sven Rahmann, Jörg Stelling et Pierre Taberlet. Ces actes contiennent les articles associés à l'ensemble de ces présentations, ainsi que la liste des quelques 130 posters qui seront affichés et discutés lors de la conférence.

C'est bien sûr avec une certaine émotion que nous avons refait vivre JOBIM à Montpellier cette année. Nous souhaitons remercier l'ensemble des membres du comité d'organisation et du comité de programme, en particulier Alban Mancheron qui a géré toute la procédure de soumission des articles et la mise en forme de ces actes, ainsi que les six conférenciers invités qui malgré des emplois du temps chargés ont accepté de présenter leur travaux pendant ces journées. Nous souhaitons une longue vie à JOBIM et le meilleur succès à ceux qui reprendront le flambeau dans les années à venir.

Gilles Caraux, Olivier Gascuel, Vincent Lefort et Marie-France Sagot

Comité d'organisation

Gilles Caraux et Vincent Lefort

Anne-Muriel Arigon-Chifolleau
Séverine Bérard
Vincent Berry
Laurent Bréhélin
Annie Château
François Chevenet
Christelle Dantec
Alexandre Dehne-Garcia
Alexis Dereeper
Jean-Baka Domelevo-Entfellner
Jean-François Dufayard
Patrice Duroux
Cécile Fleury
Philippe Gambette
Olivier Gascuel
Élisabeth Gréverie
Valentin Guignon
Laurent Journot
Matthieu Jung
Mathieu Lajoie
Pierre Larmande
Marie-Paule Lefranc
Philippe Letourmy
Alban Mancheron
Martine Marco
Isabelle Mougenot
Nicolas Philippe
Claire Poiron
Pierre Riou
Éric Rivals
Manuel Ruiz
Véronique Sals-Vettorel
Lucile Soler
Aubin Thomas
Raluca Uricaru

Comité de programme

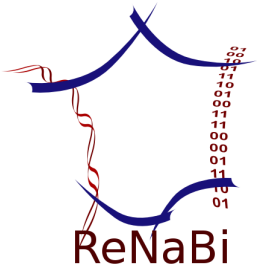
Olivier Gascuel et Marie-France Sagot

Sébastien Aubourg
Vincent Berry
Guillaume Beslon
Mathieu Blanchette
Michaël Blum
Christine Brun
Mathilde Carpentier
Giacomo Cavalli
Claudine Chaouiya
Cédric Chauve
Sarah Cohen-Boulakia
Philippe Derreumaux
Julien Dutheil
Thomas Faraut
Yann Guermeur
Andrey Kajava
Vincent Lacroix
Thierry Lecroq
Frederique Lisacek
Juliette Martin
Catherine Matias
Yves Moreau
Cédric Notredame
Pierre Peterlongo
Anne Poupon
Adrien Richard
Éric Rivals
Stéphane Robin
Marc Robinson-Rechavi
Hugues Roest-Crollius
Delphine Ropers
Manuel Ruiz
Irena Rusu
Anne Siegel
Julie Thompson
Hélène Touzet
Jacques van Helden
Yves Vandenbrouck
Jean-Philippe Vert
Louis Wehenkel

Relecteurs additionnels

Rémi Bonidal
Emmanuel Cornillot
Éric Fanchon
Guillaume Fertin
Mathieu Giraud
Stéphane Janot
Alexandra Louis
Laurent Noé
Mathieu Rouard
Fabienne Thomarat

Remerciements



Journées satellites

Lundi 6 septembre 2010

Biodiversité et Bioinformatique :
Journée organisée par Nicolas Galtier et Arnaud Estoup.
Site web : http://www.jobim2010.fr/?q=fr/node/24
Contact : Nicolas.Galtier@univ-montp2.fr

Débuter une carrière en bioinformatique :
Journée organisée par l'Association rsg-France – JeBiF.
Site web : http://www.jebif.fr/
Contact : iscb.rsg.france@gmail.com

moqa *(Méta-données et Ontologies pour la Qualité des Annotations)* :
Journée organisée par Isabelle Mougenot.
Site web : http://www.jobim2010.fr/?q=fr/node/38
Contact : isabelle.mougenot@lirmm.fr

*Modgraph*ii *(Modèles graphiques probabilistes pour l'intégration de données hétérogènes et la découverte de modèles causaux en biologie)* :
Journée organisée par Florence d'Alché-Buc, Simon de Givry, Louis Wehenkel, Philippe Leray, Gérard Ramstein et Christine Sinoquet.
Site web : http://www.lina.univ-nantes.fr/conf/modgraph2010/
Contact : modgraph@univ-nantes.fr

Vendredi 10 septembre 2010

Annotations des génomes et génomique comparée :
Journée organisée par Karyn Mégy et Stéphanie Sidibé-Boks.
Site web : http://www.jobim2010.fr/?q=fr/node/42
Contact : kmegy@ebi.ac.uk

Modélisation dynamique et simulation des réseaux biologiques :
Journée organisée par Grégory Batt, Jérémie Bourdon, Claudine Chaouiya, Hidde de Jong, Damien Eveillard, Adrien Richard, Delphine Ropers, Olivier Roux, Anne Siegel et Denis Thieffry.
Site web : http://www.jobim2010.fr/?q=fr/node/36
Contact : satellite_modelisation@sympa.univ-nantes.fr

Sommaire

Conférences invitées

Conférence invitée

Guilhem Faure, Albane Gaubert, Françoise Ochsenbein et Raphaël Guérois

céa, iBiTecS / cnrs
91191 Gif sur Yvette, France
raphael.guerois@cea.fr

Dynamic Assembly of Proteins: characterization, prediction and design

Cell processes are tightly regulated by intricate network of protein interactions. Protein interaction maps obtained for different model organisms are now providing a wealth of data to further disentangle the molecular logic associated with proteins dynamic assemblies. In particular, competitions and synergies existing between interacting partners need to be further uncovered through targeted perturbations at complex interfaces. Abrogating or perturbing specifically the edges of an interaction map remains a difficult challenge which can be bolstered through the structural description of a protein complex interface. How predictive approaches in the field of structural bioinformatics may help unravel the physical reality underlying protein interaction networks?

We are exploring how the physico-chemical properties of interfaces and the constraints arising from partners coevolution can be combined to better model the structures of protein complexes. Sequence alignments and evolutionary constraints can nowadays be successfully used to predict the 3D structure of a monomeric protein even when sequences have dramatically diverged [1-5]. How far evolutionary constraints may also be used to predict the way proteins assemble ? Coupling together computational and experimental approaches, we developed and assessed several methodologies which use sequence and structural information to better predict protein interactions [6-8]. We have particular interest in assembly chaperones, a class of proteins which regulates macromolecular assemblies and play important roles in cell stress responses. The possibility to predict how proteins do assemble, also opens perspectives for the design of compounds able to challenge native interactions achieved by these chaperones in the cellular context.

References

[1] J. Soding. Protein homology detection by hmm-hmm comparison. *Bioinformatics*, 21, pp. 951–960, 2005.

[2] B. Le Tallec, M. B. Barrault, R. Courbeyrette, R. Guérois, M. C. Marsolier-Kergoat and A. Peyroche. 20S proteasome assembly is orchestrated by two distinct pairs of chaperones in yeast and in mammals. *Mol Cell*, 27, pp. 660–674, 2007.

[3] Y. Wang, R. I. Sadreyev and N. V. Grishin. Procain: protein profile comparison with assisting information. *Nucleic Acids Res*, 37, pp. 3522–3530, 2009.

[4] A. Lopes, G. Faure, M. A. Petit and R. Guérois. Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. *Nucleic Acids Res*, in revision, 2009.

[5] B. Le Tallec, M. B. Barrault, R. Guérois, T. Carre and A. Peyroche. Hsm3/S5b participates in the assembly pathway of the 19S regulatory particle of the proteasome. *Mol Cell*, 33, pp. 389–399, 2009.

[6] H. Madaoui and R. Guérois. Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proc Natl Acad Sci* usa, 105, pp. 7708–7713, 2008.

[7] Y. Kadota, B. Amigues, L. Ducassou, H. Madaoui, F. Ochsenbein, R. Guérois and K. Shirasu. Structural and functional analysis of SGT1-HSP90 core complex required for innate immunity in plants. *EMBO Rep*, 9, pp. 1209–1215, 2008.

[8] L. Malivert, V. Ropars, M. Nunez, P. Devret, S. Miron, G. Faure, R. Guérois, J. P. Mornon, P. Revy, J. B. Charbonnier *et al.*. Delineation of the XRCC4 interacting region in the globular head domain of cernunnos/XLF. *J. Biol Chem*, 2010.

Conférence invitée

Jean-Christophe Olivo-Marin

Institut Pasteur Paris,
Unité d'Analyse d'Images Quantitative, CNRS URA 2582
25, rue du docteur Roux, 75724 Paris Cedex 15, France
jcolivo@pasteur.fr

Cells, Images and Numbers: a numerical view at biological imaging

An increasing number of biological projects aim at elucidating the links between cellular function and phenotype through imaging and modelling the spatiotemporal characteristics of cellular dynamics. This talk will present innovative methods and algorithms for the processing and quantification of $3D + t$ dynamic imaging sequences in biological microscopy and their use in biological imaging. Thanks to these tools, it is possible in a large number of experiments to automate the extraction of quantitative data from images and to facilitate the understanding of the biological information contained therein. We will present and discuss some recent developments of robust and automated tools and software for flexible and robust quantitative analysis and assessment of microscopy data. We will demonstrate algorithms for multi-particle tracking and active contours models for cell shape and motility analysis and will illustrate their application in a number of cell biology and neurosciences projects.

References

[1] M.-T. Melki, H. Saïdi, A. Dufour, J.-C. Olivo-Marin and M.-L. Gougeon. Escape of HIV-1-Infected Dendritic Cells from TRAIL-Mediated NK Cell Cytotoxicity. A pivotal Role of HMGB1. *PLoS Pathogens*, 6, 4, e1000862, 2010.

[2] S. Berlemont and J.-C. Olivo-Marin. Combining Local Filtering and Multiscale Analysis for Edge, Ridge and Curvilinear Objects Detection, IEEE *Trans. Image Processing*, 19:1, pp. 74–84, 2010.

[3] N. Chenouard, A. Dufour and J.-C. Olivo-Marin. Tracking algorithms chase down pathogens, *Biotechnology Journal*, 4:6, pp. 838–845, 2009.

[4] K. Gousset, E. Schiff, C. Langevin, Z. Marijanovic, A. Caputo, D.-T. Browman, N. Chenouard, F. de Chaumont, A. Martino, J. Enninga, J.-C. Olivo-Marin, D. Mannel and C. Zurzolo. Prions hijack tunneling nanotubes for intercellular spread. *Nature Cell Biology*, 11:3, pp. 328–336, 2009.

[5] B. Zhang, J. Zerubia and J.-C. Olivo-Marin. Gaussian approximations of fluorescence microscope point-spread function models. *Applied Optics*, 46:10, pp. 1819–1829, 2007.

[6] N. Arhel, A. Genovesio, K.-A. Kim, S. Miko, E. Perret, J.-C. Olivo-Marin, S. Shorte and P. Charneau. Quantitative four-dimensional tracking of cytoplasmic and nuclear HIV-1 complexes. *Nature Methods*, 3:10, pp. 817–824, 2006.

[7] G. Cabal, A. Genovesio, S. Rodriguez-Navarro, C. Zimmer, O. Gadal, A. Lesne, H. Buc, F. Feuerbach-Fournier, J.-C. Olivo-Marin, E.-C. Hurt and U. Nehrbass. SAGA interacting factors confine sub-diffusion of transcribed genes to the nuclear envelope.*Nature*, 441, pp. 770–773, 2006.

[8] A. Genovesio, T. Liedl, V. Emiliani, W. Parak, M. Coppey-Moisan and J.-C. Olivo-Marin. Multiple particle tracking in $3D + t$ microscopy: method and application to the tracking of endocytozed Quantum Dots. IEEE *Trans. Image Processing*, 15:5, pp. 1062–1070, 2006.

Conférence invitée

Luis Quintana-Murci

Institut Pasteur Paris,
CNRS URA3012
25, rue du docteur Roux, 75724 Paris Cedex 15, Paris, France
quintana@pasteur.fr

Human Genome Diversity: from demography to natural selection

Different environmental, demographic and selective forces, together with cultural and social characteristics of human lifestyle, shape the patterns of variability of the human genome at the population level. A detailed description of the relative weight and influence of these processes, which may vary among individuals and populations, will provide important insights into human evolutionary history, which might, in turn, also facilitate identification of complex disease genes. Our research activities cover two highly inter-related areas: the study of genetic diversity at noncoding regions of the genome, from which we can infer historical and demographic parameters characterizing human populations, and the study of diversity in genomic regions involved in immune response or host-pathogen interactions, with which we can unmask the footprints of natural selection exerted by pathogens on the host genome. I will review our most recent data on these different aspects, by focusing on specific examples of demography, lifestyle and natural selection. These include: (i) the influence of modes of subsistence and lifestyle – the transition from hunter-gathering to farming – on the demographic and adaptive history of human populations, by focusing on the case of Pygmy hunter-gatherers and neighbouring farmers from Central Africa. (ii) The study of how natural selection has targeted human miRNAs as a model system for investigating the influence of natural selection on gene regulation. Indeed, more than 30% of human genes are thought to be regulated by miRNAs, and their role in diverse physiological processes, including development, growth, differentiation and metabolism is increasingly recognized. (iii) The value of the evolutionary and population genetics approach in the context of host-pathogen interactions. Detecting and identifying the extent and type of natural selection acting on genes involved in immunity-related processes provide insights into immunological host defense mechanisms and highlight pathways playing an important role in pathogen resistance. Altogether these studies, based on a multi-locus approach and considering the different forces shaping the patterns of human genome variability, shed light onto the complex demographic and adaptive history of our species.

Conférence invitée

Sven Rahmann

TU Dortmund
Bioinformatics for High-Throughput Technologies
Computer Science 11
TU Dortmund
44221 Dortmund, Germany
Sven.Rahmann@uni-dortmund.de

Algorithmic Challenges from New Sequencing Technologies

New high-throughput sequencing technologies are able (and will be even more so in the future) to produce sequence data at a higher rate than present methods can analyze it. At the same time, applications are quite diverse: de novo sequencing and re-sequencing of genomes, SNP discovery, determination of methlylation state, classical gene expression analysis by mRNA (or tag) sequencing, short RNA expression analysis, ChIP-seq, just to name a few.

In my talk, I will present analyses of different datasets from different sequencing technologies that we conducted in collaboration with two groups from the University Hospital Essen: microRNA expression in favorable and unfavorable neuroblastoma subtypes [1] (with the Pediatric Oncology Department), and methylation state of CpG islands in human blood and sperm cells [2] (with the Human Genetics Department). In particular, I will highlight the challenges we faced beyond the standard read mapping procedure.

Next, I will discuss the implications of the new sequencing technologies for phylogenetic analyses and argue that novel ideas for explorative analysis of multiple sequence alignments are needed. I will present one idea developed in collaboration with the Bioinformatics group at the University of Würzburg [3].

Finally, I will present my vision about probabilistic models to efficiently describe large pan-genomes (that will result from 1000-genome projects, for example), and my opinion on the required research to develop such models (beyond sheer computing power).

References

[1] J. H. Schulte, T. Marschall, M. Martin, P. Rosenstiel, P. Mestdagh, S. Schlierf, T. Thor, J. Vandesompele, A. Eggert, S. Schreiber, S. Rahmann and A. Schramm. Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma. *Nucl. Acids Res.*, 2010.

[2] M. Zeschnigk, M. Martin, G. Betzl, A. Kalbe, C. Sirsch, K. Buiting, S. Gross, E. Fritzilas, B. Frey, S. Rahmann and B. Horsthemke. Massive parallel bisulfite sequencing of CG-rich DNA fragments reveals that methylation of many X-chromosomal CpG islands in female blood DNA is incomplete. *Human Molecular Genetics*, 18:8, pp. 1439–1448, 2009.

[3] R. Schwarz, P. N. Seibel, S. Rahmann, C. Schön, M. Hünerberg, C. Müller-Reible, T. Dandekar, R. Karchin, J. Schultz and T. Müller. Detecting species-site dependencies in large multiple sequence alignments. *Nucl. Acids Res.*, 7:18, pp. 5959–5968, 2009.

Conférence invitée

Jörg STELLING

Swiss Federal Institute of Technology Zürich,
Department of Biosystems Science and Engineering
1058 8.00 Mattenstrasse 26, 4058 Basel, Switzerland
joerg.stelling@bsse.ethz.ch

Computational Engineering of Synthetic Gene Circuits

Ultimately, synthetic biology aims at establishing novel, useful biological functions by suitably combining well-characterized parts. Especially when complex circuits – in terms of the number of components and interactions involved, or with respect to the dynamic behavior – are to be designed, computational engineering methods have to be an integral part of the approach. Here, we will focus on engineering concepts to achieve scalability and robustness (relative insensitivity to external or internal perturbations) of designed circuits. Both aspects are important for the field because the biology-based parts employed are not (yet) well-characterized, the circuits have to operate in a noisy cellular environment, and they cannot be completely isolated from the cellular context. Specific examples that illustrate the challenges of and possible strategies for rational circuit design include devices for time-delayed gene expression, tunable synthetic oscillators, and physiological set-point controllers in mammalian cells. These cases demonstrate that both novel mathematical modeling and systems analysis methods are needed to enable efficient computational design of synthetic circuits.

Conférence invitée

Éric Coissac et Pierre Taberlet

Laboratoire d'Écologie Alpine, cnrs umr 5553
Université Joseph Fourier, bp 53, 38041 Grenoble Cedex 9, France
pierre.taberlet@ujf-grenoble.fr

Biodiversity and dna Barcoding

Dna barcoding – taxon identification using a standardized dna – is mainly developed through an international initiative (Consortium for the Barcode of Life, CBoL, http://barcoding.si.edu). Dna barcoding *sensu stricto* corresponds to the identification of one specimen to the species level using a single standardized dna fragment. This definition fits with the CBoL view. Dna barcoding *sensu lato* corresponds to the identification of a set of organisms present in an environmental sample to any taxonomical level using any dna fragment (dna metabarcoding). Our scientific objective is to use the metabarcoding approach to analyze biodiversity using environmental samples (water, soil, etc.). The experimental protocol consists (i) to sample in the field, (ii) to extract dna, (iii) to amplify dna using universal primers, (iv) to sequence the pcr product using next generation sequencers (454, Solexa), and (v) to assign the sequences to the relevant taxa. We currently focus on plants and animals. Such an approach represents real challenges at the bioinformatic level, both before carrying out the experiments, and after obtaining the output files of the sequencer. How to design optimal primers? How to test *in silico* these primers for specificity and accuracy? How to design an efficient tagging system for being able to properly assign a sequence read to a sample in a sequencing experiments where hundred of samples were mixed together? How to assign sequences to a taxon, with or without reference sequences? How to deal with amplification/sequencing errors? We will present the different bioinformatic tools especially developed for analyzing environmental samples using the dna barcoding concept (http://www.grenoble.prabi.fr/trac.OBITools). Then, we will show some results concerning the identification of plants and animals from soil samples, or from feces for diet analysis.

Présentations longues

The carbon assimilation network in *Escherichia coli* is densely connected and largely sign-determined by directions of metabolic fluxes

Valentina BALDAZZI[1,4], Delphine ROPERS[1], Yves MARKOWICZ[1,2], Daniel KAHN[1,3], Johannes GEISELMANN[1,2] and Hidde DE JONG[1]

[1] INRIA Grenoble - Rhône-Alpes, 655 avenue de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France
{Delphine.Ropers,Hidde.de-Jong}@inria.fr
[2] Laboratoire Adaptation et Pathogénie des Microorganismes, UMR 5163 CNRS, Université Joseph Fourier, Bâtiment Jean Roget, Faculté de Médecine-Pharmacie, Domaine de la Merci, 38700 La Tronche, France
{yves.markowicz,hans.geiselmann}@ujf-grenoble.fr
[3] Laboratoire de Biométrie et Biologie Evolutive, UMR 5558 CNRS, Université Lyon 1, INRA, 43 bd du 11 novembre 1918, 69622 Villeurbanne, France
kahn@biomserv.univ-lyon1.fr
[4] INRA, Unité Plantes et SystÃmes de culture Horticoles, Domaine St Paul, Agroparc, 84941 Avignon Cedex 9, France
valentina.baldazzi@avignon.inra.fr

Abstract *Gene regulatory networks consist of direct interactions, but also include indirect interactions mediated by metabolites and signaling molecules. We describe how these indirect interactions can be derived from a model of the underlying biochemical reaction network, using weak time-scale assumptions in combination with sensitivity criteria from metabolic control analysis. We apply this approach to a model of the carbon assimilation network in Escherichia coli. Our results show that the derived gene regulatory network is densely connected, contrary to what is usually assumed. Moreover, we show that the signs of the indirect interactions are largely fixed by the direction of metabolic fluxes, independently of specific parameter values and rate laws, and that a change in flux direction may invert the sign of indirect interactions. This leads to a feedback structure that is at the same time robust to changes in the kinetic properties of enzymes and that has the flexibility to accommodate radical changes in the environment.*

Keywords System biology, gene regulatory network, metabolism, *E.coli.*

The adaptation of bacteria to changes in their environment involves adjustments in the expression of genes coding for enzymes, regulators, membrane transporters, *etc.* [1,2,3]. These adjustments are controlled by gene regulatory networks ensuring the coordinated expression of clusters of functionally related genes. The interactions in the network may be direct, as in the case of a gene coding for a transcription factor regulating the expression of another gene. Most of the time, however, regulatory interactions are indirect, *e.g.* when a gene encodes an enzyme producing a transcriptional effector [4].

A gene regulatory network can thus not be reduced to its transcriptional regulatory interactions: by ignoring indirect interactions mediated by metabolic and signaling pathways we may miss crucial feedback loops in the system. The network controlling carbon uptake in the bacterium *Escherichia coli* is a good example because it integrates metabolism, signal transduction, and gene expression. At the level of gene expression, the network includes intricate feedback loops that arise from indirect interactions between the subsystems. Global regulators like Crp control expression of enzymes in carbon metabolism [5,6], while intermediates of the latter pathways control the expression of global regulators. For instance, the phosphorylation of EIIA activates adenylate cyclase (Cya) to produce cAMP which is required for the activation of Crp [7,8].

The aim of this paper is to develop a method for the systematic derivation of direct and indirect interactions in a gene regulatory network from the underlying biochemical reaction network. Due to the complexity of the intermediate metabolic and signaling networks, determining indirect interactions is difficult in general. We show that model reduction based on weak assumptions on time-scale hierarchies in the system [9,10,11], together with sensitivity criteria from metabolic control analysis [10,12], are able to uncover such interactions. Our approach starts from a model of the biochemical reaction system in the form of a system of or-

dinary differential equations. We reformulate this system into coupled fast and slow subsystems, by distinguishing between reactions that are fast and slow in the physiological range of interest, and by redefining fast and slow variables accordingly. This is rather straightforward to achieve for the types of systems considered here, as enzymatic and complex formation reactions are typically fast on the time-scale of protein synthesis and degradation. Assuming that the fast subsystem is at quasi-steady state, the indirect interactions between genes are now defined by the Jacobian matrix of the slow system. The advantage of this approach is that it does not require fully specified kinetic models with known rate laws for reaction rates and numerical values for the parameters: the dependencies of the reaction rates on metabolite and enzyme concentrations are usually sufficient once the metabolic flux directions are fixed.

We apply our method to a model of the upper part of the carbon assimilation network in *E. coli*, consisting of the glycolysis and gluconeogenesis pathways and their genetic and metabolic regulation. The analysis of the derived gene regulatory network leads to three new insights. First, contrary to what is often assumed, the network is densely connected due to numerous feedback loops resulting from indirect interactions. This additional complexity is operative on the time-scale of gene expression and represents an important issue for the correct interpretation of data from genome-wide transcriptome studies. For instance, our method correctly predicts that a *pykF* deletion leads to increased expression of *fruR* and decreased expression of *cya* during glycolysis [13]. The second and most remarkable conclusion of our study of the *E. coli* network is that for given growth conditions, the signs of the indirect interactions are largely independent of the exact form of kinetic rate laws and precise parameter values. Therefore the feedback structure is robust to changes in kinetic properties of enzymes and other biochemical reactions species. However a radical changes in the environment, *e.g.*, the exhaustion of glucose, may invert the signs of fluxes, and thus of indirect interactions, resulting in a dynamic rewiring of the regulatory network. Such an overall modification of the control architecture in response to environmental perturbations may be beneficial to the cell, as it increases its adaptive flexibility.

More details on this work can be found in a recent publication in *PloS Computational Biology* [14].

References

[1] Faith J, Hayete B, Thaden J, Mogno I, Wierzbowski J, et al. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. PLoS Biol 5: 0054-0066.

[2] Oh M, Rohlin L, Kao K, Liao J (2002) Global expression profiling of acetate-grown *Escherichia coli*. J Biol Chem 277: 13175-13183.

[3] Friedman N, Vardi S, Ronen M, Alon U, Stavans J (2005) Precise temporal modulation in the response of the SOS DNA repair network in individual bacteria. PLoS Biol 3: e238.

[4] Brazhnik P, de la Fuente A, Mendes P (2002) Gene networks: How to put the function in genomics. Trends Biotechnol 20: 467-472.

[5] Gutierrez-Ríos R, Freyre-Gonzalez J, Resendis O, Collado-Vides J, Saier M, et al. (2007) Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli*. BMC Microbiol 7: 53.

[6] Hardiman T, Lemuth K, Keller MA, Reuss M, Siemann-Herzberg M (2007) Topology of the global regulatory network of carbon limitation in *Escherichia coli*. J Biotechnol 132: 359-374.

[7] Park YH, Lee B, Seok YJ, Peterkofsky A (2006) *In vitro* reconstitution of catabolite repression in *Escherichia coli*. J Biol Chem 281: 6448-6454.

[8] Saier MJ, Ramseier Jr T (1996) Regulation of carbon utilization. In: Neidhardt F, Curtiss III R, Ingraham J, Lin E, Low K, et al., editors, *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, Washington D.C.: ASM Press. pp. 1325-1343.

[9] Jamshidi N, Palsson BO (2008) Formulating genome-scale kinetic models in the post-genome era. Mol Syst Biol 4: 171.

[10] Heinrich R, Schuster S (1996) The Regulation of Cellular Systems. New York: Chapman & Hall.

[11] Okino M, Mavrovouniotis M (1998) Simplification of mathematical models of chemical reaction systems. Chem Rev 98: 391-408.

[12] Fell D (1996) Understanding the control of metabolism. Portland Press.

[13] Siddiquee K, Araúzo-Bravo, M K Shimizu (2004) Effect of a pyruvate kinase (*pyk*F gene) knockout mutation on the control of gene expression and metabolic fluxes in *Escherichia coli*. FEMS Microbiol Lett 235: 25-33.

[14] Baldazzi V, Ropers D, Markowicz Y, Kahn D, Geiselmann J, de Jong H (2020) The carbon assimilation network in *Escherichia coli* is densely connected and largely sign-determined by directions of metabolic fluxes. PLoS Comput Biol 6: e1000812.

Design and exploitation of a versatile Arabidopsis whole-Genome Tiling Array

Tiling-array data: analysis and visualization

Caroline BÉRARD[1], Sandra DEROZIER[2], Sandrine BALZERGUE[2], Tristan MARY-HUARD[1], François ROUDIER[3], Stéphane ROBIN[1], Alain LECHARNY[2], Vincent COLOT[3], Michel CABOCHE[2], Sébastien AUBOURG[2] and Marie-Laure MARTIN-MAGNIETTE[1,2]

[1] UMR AgroParisTech/INRA MIA 518, 16 rue Claude Bernard, 75231 Paris Cedex 05, France.
[2] Unité de Recherche en Génomique Végétale (URGV), UMR INRA 1185-UEVE ERL CNRS 8196, 2 rue Gaston Crémieux, CP 5708, Evry, France.
[3] Institut de Biologie de l'ENS (IBENS), CNRS UMR8197 - INSERM U1024, 46 rue d'Ulm 75230 Paris, France.
caroline.berard@agroparistech.fr, derozier@evry.inra.fr

Keywords Tiling-array, statistics, database, plant genomics, data integration

1 Introduction

The ANR/Génoplante Tiling Array Genome (TAG) project aims to design, validate and exploit a unique chip covering the *Arabidopsis thaliana* genome. The applications of this chip concern various types of experiments such as ChIP-chip to study DNA methylation and histone modifications or transcriptome experiments to detect and analyze coding and non-coding transcripts. For each type of data, we have developed an adapted statistical method and a visualization tool of the results in the FLAGdb++ environment.

2 Tiling-array features

The nuclear, plastidial and mitochondrial Arabidopsis chromosomes have been segmented in 160 bp-long regions in which oligonucleotides have been designed by the NimbleGen Company. Oligonucleotides (*i.e.* probes) have sizes ranged from 50 to 75 mers and are selected in order to have a near constant Tm (around 76˚C). Beyond the chromosomal position, the hybridization quality is the priority criteria for the design. The whole Arabidopsis genome is finally represented by 1.4 million of probes, 717 246 covering each strand. The repeat sequences of the Arabidopsis genome (mainly recently duplicated genes or transposable elements) results in the presence of 4.5% of not specific probes for which cross-hybridization is probable. The resulting array is composed of 3 hybridization rooms in which the 717 246 probes have been synthesized *in situ* by a photolithographic process.

3 Statistical Methods

The methods developed were first used to analyze data from *Arabidopsis thaliana* but are not organism-dependent. R packages are available.

3.1 ChIPmix [1]: Analysis of ChIP-chip data

In a two-color ChIP-chip experiment, two samples are compared: DNA fragments crosslinked to a protein of interest (IP), and genomic DNA (Input). The IP signal depends not only on the status (normal/enriched) of the probe, but also on the INPUT signal. This dependence is not taken into account by working with the usual ratio IP/INPUT. For this reason we directly consider the two intensities log-IP and log-INPUT, denoted (x_i, Y_i) for the probe i, respectively. The (unknown) status of the probe is characterized through a label Z_i which equals 1 if the probe is enriched and 0 if it is normal (not enriched). We assume the Input-IP relationship to be linear whatever the population, but with different slope and intercept. More precisely, we have:

$$Y_i = a_0 + b_0 x_i + \mathcal{N}(0, \sigma^2) \text{ si } Z_i = 0,$$
$$Y_i = a_1 + b_1 x_i + \mathcal{N}(0, \sigma^2) \text{ si } Z_i = 1.$$

We control the probability for a probe to be wrongly assigned to the enriched class. This control is similar to the one used in the hypothesis test theory. We applied this method on several ChIP-chip data to study the impact of histone modifications on the regulation of gene expression.

3.2 Bidimensionnal Gaussian Mixture: Analysis of transcriptome data

The NimbleGen tiling array is also used to compare two transcript samples without *a priori* on the transcript regions. We expect to distinguish four different groups (cf Fig. 1): a group of non-hybridized probes, a group where gene expression is identical in the two conditions, and two groups in which gene expression differs between the two conditions. We propose to model these data with a two-dimensional Gaussian mixture with four components with constraints of the variance matrix [2] to take biological knowledge into account. The model originality is to consider both the dependency between neighboring probes with a HMM model and also the known probe annotation. We know that a probe behaves differently if it covers an exon, an intron or an intergenic region. First application concerns a comparison between the leaf and the seed.

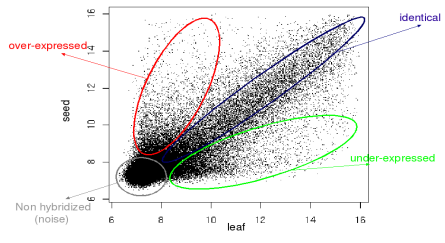


Fig. 1. Schematic representation of the 4 groups of probes

4 Visualization and exploitation of tiling-array data

Visualization of data is central in genomics. It is the first step in hypothesis inferences from versatile tools delivering exhaustive information on genome transcription. Visualization of several types of information suggests links between genome characteristics and expression. Some unexpected predictions may, after experimental validation, provide new knowledge[3].

4.1 FLAGdb++

FLAGdb++ (http://urgv.evry.inra.fr/FLAGdb++) is an integrative database dedicated to the structural and functional genomics of plants [4]. Up to now, FLAGdb++ is mainly focused on the sequenced genomes of *Arabidopsis thaliana*, Oryza sativa, Vitis vinifera and Populus trichocarpa. It helps biologists to study the function of plant genes in considering them in a wide context : a multigene family, a topological environment, and/or a functional network. FLAGdb++ is composed of a relational database and an associated user-friendly interface (in Java). Different tools have been developed with a conceptual effort for the graphical display and the hierarchical organization of the different genomic data to browse and explore them and decipher functional relationships between them.

4.2 A new tiling-array module

A Java module has been developed and added to FLAGdb++ to display the genome-wide data produced by tiling chips. The visualization concerns the probe features (sequences, Tm, uniqueness in genome, cross-hybridization risk with paralogous genes), their position relative to the structural annotation and, the statistical hybridization results. Since data quantity is huge, a Derby database for the client side is used to provide a flowing display of results on large genomic regions. Through a project management tool, the users have the possibility to import private results to analyse them in the full data background proposed by FLAGdb++. The interfaces help the genome exploration and reveal unexpected transcriptional activities in highlighting new RNA genes, regulatory antisens transcripts and alternative splicing events. Their fine characterization at the genome scale is currently running.

Acknowledgements

This work was supported by the ANR/Genoplante project TAG. C. Bérard's PhD is funded by MIA, GAP and MICA departments of INRA.

References

[1] M-L. Martin-Magniette, T. Mary-Huard, C. Berard and S. Robin, ChIPmix: mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics*, 24:i181-i186, 2008.

[2] C. Bérard, M-L. Martin-Magniette, A. To, F. Roudier, V. Colot, and S. Robin, Mélanges gaussiens bidimensionnels pour la comparaison de deux échantillons de chromatine immunoprécipitée. *La revue de MODULAD*, 40:53-68, 2009.

[3] S. Aubourg, M-L. Martin-Magniette, V. Brunaud, L. Taconnat, F. Bitton, S. Balzergue, P-E. Jullien, M. Ingouff, V. Thareau, T. Schiex, A. Lecharny, J-P. Renou, Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. *BMC Genomics*, 8: 401, 2007.

[4] F. Samson, V. Brunaud, S. Duchêne, Y. De Oliveira, M. Caboche, A. Lecharny, S. Aubourg, FLAGdb++: a database for the functional analysis of the Arabidopsis genome. *Nucleic Acids Research*, 32: D347-D350, 2004.

Mining microarray data for regulatory interactions with TranscriptomeBrowser.

Aurélie BERGON[1,2,*], Cyrille LEPOIVRE[1,2*], Fabrice LOPEZ[1,2,*], Denis THIEFFRY[2,3], Jean IMBERT[1,2], Christine BRUN[1,2], Carl HERRMANN[1,2] and Denis PUTHIER[1,2]

[1] Université de la Méditerranée, Marseille, France
[2] TAGC (INSERM U928), Marseille, France
{bergon,lepoivre,lopez,herrmann,imbert,brun,puthier}@tagc.univ-mrs.fr
[3] IBENS - CNRS UMR 8197 / INSERM U1024, Ecole Normale Supérieure, Paris, France
thieffry@ens.fr
[*] These authors contributed equally to this work and are listed in alphabetical order.

While microarray data are rapidly accumulating in public repositories, novel methods are needed to mine and visualise OMICS data. As genes are part of complex regulatory networks, typical data mining software focus on capturing clusters of co-expressed genes from microarray data. The TranscriptomeBrowser [1] (TBrowser) project, aims at easing access to public microarray experiments. Its specificities rely on (i) the systematic annotation of automatically extracted expression signatures and (ii) the development of a powerful search engine to mine relevant informations. TBrowser software is extremely useful (i) to quickly re-analyse any experiment stored in the database (ii) to search for biological contexts in which a set of genes are co-expressed or in which any biological function is highly represented (ii) and to compare expression signatures obtained through different experiments.

Starting with a set of 1,400 experiment in the first release [1], we have conducted a new analysis on 3,000 experiments derived from 105 microarray platforms (51 species). Unsupervised classification using the "Density Based Filtering and Markov CLustering" algorithm (DBF-MCL) [1] led to the extraction of 30,000 expression signatures. We further performed a systematic functional annotation of the resulting expression signatures using a large compendium of terms derived from gene annotation databases (*eg;* WikiPathway, REACTOME, HMDB, GeneSigDB,...) and meta-databases (DAVID knowledgeBase). Importantly, the new release of TBrowser also included numerous informations related to cis-regulatory motifs that may be involved in transcriptional (*e.g*; ECRBase, CisRED, TFBSConserved, LymphTF-DB, OREGANNO) and post-transcriptional (*e.g;* TargetScan, Pictar) control of gene expression (currently 88,953 gene to motifs relationships). Altogether we have have currently stored 541,516 keywords and 40,659,707 gene to term relationship. Interestingly, our strategy of systematic annotation of transcriptomic data identified numerous putative regulatory events, whose relevance is strengthened by biological process enrichment analysis. For example, expression signatures enriched in E2F targets are also enriched for GO term "Cell cycle", enrichment for MEF2A targets are found in signatures enriched for "muscle contraction" term, RFX1 targets enrichment is related to "ciliary or flagellar motility", while IRF1 targets are observed in signature related to "immune response". All these informations can be accessed through the Java application. Annotation terms from various databases can be combined to focus on the most relevant expression signatures. For example, genes related to unfolded protein Response (UPR) can be found by searching for expression signatures enriched for predicted XBP1 targets and for genes related to "endoplasmic reticulum unfolded protein response" (GO term). Although some obvious transcription factor networks (e.g., co-occurence of E2F and Myc/Max target enrichments) clearly appear in the database, the flexibility of TBrowser allow to further look for alternative combinations.

When constructing of a biological network (that can be ultimately use for dynamic modelling) the first step is the definition of the system boundaries. This means that one has to define the genes that are part of the system together with their interactions. This information are most generally derived from scientific literature or obtained through re-analysis of high-throughput genomic data. As we have conducted a systematic analysis over thousand of experiments we have also captured some stricking expression signatures that underscore numerous biological processes. In order to translate any expression signature into a map highlighting putative physical and regulatory interactions, we have developed a new plugin for TBrowser, called InteractomeBrowser. InteractomeBrowser is based on the Prefuse Java library and uses a set of high level terms of the Cellular Component ontology (GO slim) to map gene products onto a schematic view of cell compartments. Predicted regulatory interactions derived from

expression signature annotations can be represented together with putative physical interactions (obtained from Intact and HPRD) and enzyme-Substrate Relationships (KEA database). As all tools and information are interconnected into an extensible and unified data mining suite, an expression signature can be easily translated into a model highlighting putative molecular events occurring in a given biological context. Results obtained using the InteractomeBrowser plugin can be further export to Cytoscape or GINsim [2,3] for further analysis or to generate dynamical models. Tbrowser and its plugins arguably constitute a very efficient way to mine public microarray repositories for biologically meaningful information.

Acknowledgements

This work is supported by EU FP7 (APO-SYS project) and ANR (SYSCOM CALAMAR project).

References

[1] F. Lopez, J. Textoris , A. Bergon, G. Didier, E. Remy, S. Granjeaud, J. Imbert, C. Nguyen, D. Puthier. TranscriptomeBrowser: a powerful and flexible toolbox to explore productively the transcriptional landscape of the Gene Expression Omnibus database. PLoS One, 3(12), 2008.

[2] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage , N. Amin, B. Schwikowski, T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res., 13(11), 2003

[3] A. Naldi, D. Berenguier, A. Fauré, A.; Lopez, F., Thieffry, D. & Chaouiya, C. Logical modelling of regulatory networks with GINsim 2.3. Biosystems, 97, 134-139, 2009.

Integrating *omics* data by using a gene neighboring based distance

Philippe BORDRON[1], Damien EVEILLARD[1] and Irena RUSU[1]

Computational Biology group (ComBi), LINA, Université de Nantes, CNRS UMR 6241, 2 rue de la Houssinière, 44300 Nantes, France
philippe.bordron@univ-nantes.fr

Abstract *As living systems are abstracted by information of different nature, in particular genomic and metabolic, their integration and interpretation into a unique framework remains challenging. We propose such a dedicated framework that integrate information that is not superposable in an obvious way. It gives rise the opportunity to investigate the impact of a given biological property like herein the genomic distance based on gene neighboring, in either a genomic or a metabolic context. In particular, we show that (1) in metabolic networks, the reaction chains (or paths) which join two given reactions diverge far less than in general networks, thus acting like a unique* supertrack *rather than like several different ways to join the two given reactions; and (2) these supertracks often represent the projection (in a sense explained in the paper) of operons on the metabolic network. Consequently, integrating metabolic and genomic knowledge using our method allows to find associations between genes, enzymes and metabolic paths involved together in the bacterial system behavior.*

Keywords systems biology, genome, gene neighboring, metabolic network, operons.

1 Introduction

A living system is abstracted by information of different nature. Their integration is an unavoidable, though only incipient, approach to identify modules which are consistent with all data sources and thus with the system. Unfortunately, information of different natures is not superposable in an obvious way, and specific approaches are usually developed for specific types of information. However, as the need to combine heterogeneous information and to analyze it as a whole is constantly increasing and diversifying, generic methods to gather informations together and to represent them in a suitable way for exploration become necessary.

In the prokaryote system, the metabolic network and the genome organization are well studied aspects. The way to structure and analyze them is the key concept of adjacency [1] or, more generaly, the concept of connectivity. This concept is easily handleable in networks, enables to associate successive adjacency information, and deals with potential missing information (edges or arcs) from the network. The integration of different types of information should therefore be based on a formalization of these concepts, as first suggested in [2]. Moreover, whereas graph theory offers both the degree of abstraction and an important part of the tools needed by such an approach, not every graph-theoretical defined notion of a neighborhood is biologically relevant [3]. It is thus important to define appropriate ways to deal with proximity across heterogeneous data, depending on the nature of the information at hand. We then aim at :

(1) integrating neighborhood/non-neighborhood information about genes (issued from sequence/structure analysis) into metabolic networks (*via* the enzymes, products of genes that catalyze the reactions),
(2) extending the local, neighborhood-based analysis to a larger scale, connectivity-based analysis, that is made possible by the use of networks,
(3) proposing a generic handling of the proximity information between genes *via* a notion of distance.

As a concrete application, our main goal is studying the concomitant (relative) proximity of metabolic genes (a) according to the aforementioned distance, and (b) in the metabolic networks (*via* the enzymes, products of genes), with the aim of finding correlations between genes, enzymes and metabolic pathways, which would explain the role and the place of each of them in the light of the bacterial system behavior. The originality of our approach stands (i) in the use of a distance instead of a binary neighbor/non-neighbor relation, (ii) in its genericity due to the potential use of *any* distance (including one that combines

information from several sources) and (iii) in the possibility to explore alternative paths (defined below using the distance). For these reasons, our approach generalizes previous works aiming at integrating, for instance, genomic and metabolic data [2,4], or genomic, coexpression and metabolic data [5,6,7,8]

Note that, in contrast of general methods that integrate heterogeneous information, our approach is independent of the distance, but the interpretation of the results is not. To perform a complete explanation of our framework, we therefore choose to illustrate it on a particular distance between genes, the *genomic distance* defined as the number of intermediate genes along the genome between two given genes, plus 1. This choice relies on the commonly accepted observation that the gene order in bacterial genomes is far from random [9,10].

This paper thus proposes an approach that integrates metabolic and bacterial genome information using *Escherichia coli* as a concrete benchmark (Sec. 2). The resulting integrated model shows the existence within the metabolic network of precisely defined substructures called *supertracks*, which are the projections on the metabolic network of many operons (Sec 3). These structures are discussed in the context of previous works (Sec. 4) and the Conclusion (Sec. 5) follows.

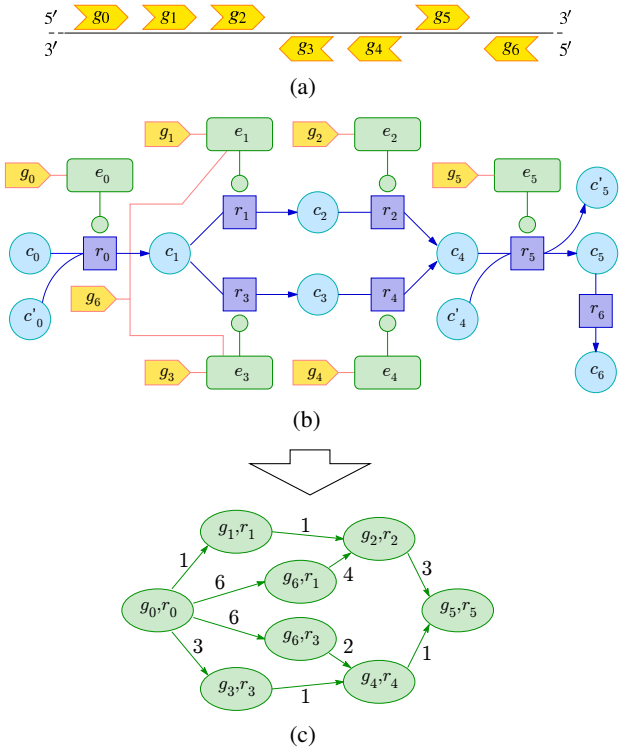


Fig. 1. Integration of genomic and metabolic information into $\mathcal{G}_{int}$. (a) A bacterial monochromosomic genome is a linear or circular sequence of genes (fat arrows). (b) Its corresponding metabolic network (in SGBN standard): compounds (circles) are substrates and/or products of reactions (squares), that are catalyzed by enzymes (rounded boxes) produced by genes (fat arrows). (c) The resulting integrated network $\mathcal{G}_{int}$, where each arc is weighted by the genomic distance between the two genes in its endpoints.

2 Material and Methods

2.1 Data

Escherichia coli (K12 MG1655, [11]) is one of the most investigated bacterial species, and quickly appeared as an accurate benchmark for our approach. At the time of the study, a set of 4 242 genes composes its circular monochromosomic genome (NCBI/GenBank). Its corresponding metabolism is composed of a set of 2 971 biochemical compounds involved in 1 131 reactions catalyzed by 647 enzymes (KEGG PATHWAYS, [12]). Among them, 558 are encoded by identified genes, as indicated by NCBI/GenBank. Each reaction is reconstructed from its association with compounds in each *E. coli* pathways map. In this manner, metabolic network is cleaned in term of hub metabolites [13] (i.e. each selected metabolite participates in less than 25 reactions). For obvious technical reasons, only the reactions catalyzed by these enzymes are concerned.

We complete our study by performing random shuffling experiments independently on the genome (by randomly modifying the gene order) and on the metabolic network (according to [14], by randomly shuffling the endpoints of the arcs in a compound-free representation of the metabolic network called representation graph). Note that such an approach might decrease the number of arcs in the network, and thus some vertex degrees.

2.2 Integrating genomic distance and metabolic network

A bacterial genome is represented as a linear or circular sequence of genes (see Fig. 1(a)). The genomic distance between two different given genes on a linear genome is the number of intermediate genes between the two genes along the genome, plus 1. A gene has a null distance between it to itself. In a circular genome, the distance consists of the minimum one obtained from the right-hand and left-hand traversal.

The relationship between a genome and its corresponding metabolic network takes place with the "gene produces enzyme(s)" rule, as illustrated in Fig. 1(b). Combination of this rule and knowledge at disposal of *E. coli* conduces to define an integrated genomic metabolic network, denoted $\mathcal{G}_{int}$ (see Fig. 1(c) for illustration). The network $\mathcal{G}_{int}$ is a directed graph,

whose vertices are all the pairs (gene g, reaction r) such that the gene g produces an enzyme (identified herein by its EC number) that catalyzes the reaction r. An arc goes from vertex (g_1, r_1) to vertex (g_2, r_2) whenever a product of r_1 is a substrate of r_2. Its weight w is defined as the genomic distance between g_1 and g_2. In a general manner, the weight of a subnetwork of $\mathcal{G}_{int}$ is the total weight of its arcs, and its *neighboring coefficient* $\overline{w}$ is the ratio between its total weight and the number of its distinct reactions (found in its vertices). Intuitively, the neighboring coefficient measures the average genomic distance between two genes involved in successive reactions from the metabolic network. A small $\overline{w}$ expresses a real gene proximity along the genome. Note herein that two successive reactions catalyzed by the same gene, produce a weight of 0, which implies that $\overline{w}$ might be less than 1.

The resulting $\mathcal{G}_{int}$ of *E. coli* is composed of 2 343 vertices and 13 288 arcs, which correspond to 1 049 metabolic reactions (92.75% of the *E. coli* reactions) and 779 genes (18.36% of the bacterial genome). Shuffling the genome conserves the structure of $\mathcal{G}_{int}$. On the contrary, shuffling the metabolic network yields a network $\mathcal{G}_{int}$ with the same number of vertices but only 11 820 arcs on average (over the 10 performed experiments).

2.3 Integrated pathways

A (directed) path from $\mathcal{G}_{int}$ with a small value for $\overline{w}$ represents a reaction chain from the metabolic network whose genes encoding for the reactions of interest are close to each other along the genome. Finding these paths reverts therefore a particular interest from a functional viewpoint, since the involved genes are concerned by the hypothesis that the gene neighboring could imply a metabolic feature.

Given any pair of reactions involved in the vertices of $\mathcal{G}_{int}$, that we identify as the source reaction and the destination reaction, we are thus interested in the (directed) paths in $\mathcal{G}_{int}$ that start with a vertex containing the source reaction and end with a vertex containing the destination reaction. The path in $\mathcal{G}_{int}$ which has the smallest $\overline{w}$, i.e. neighboring coefficient, is called the *1-Integrated Pathway* (or *1-IP*) of the two reactions. When projected in the bacterial metabolic network, this path represents the way to join two given reactions while preserving the minimum genomic distance along the genome. Fig. 2(a) shows an example of 1-IP. Similarly, for a fixed positive integer k, the subnetwork of $\mathcal{G}_{int}$ obtained as the union of the k distinct paths with smallest $\overline{w}$ joining two given reactions is called the *k-Integrated Pathway* (or *k-IP*) of the two reactions. When projected in the metabolic context,

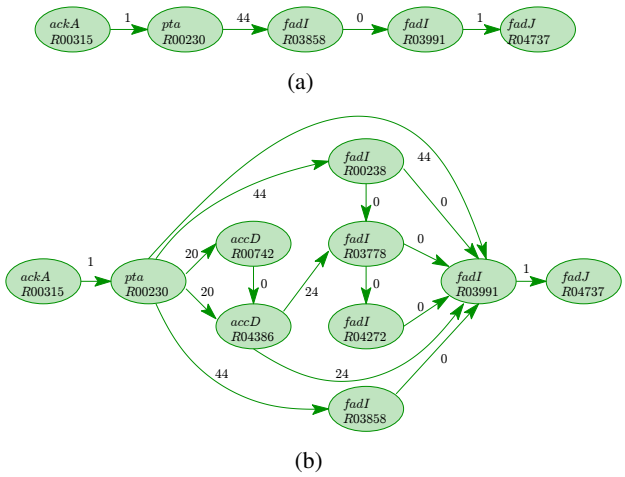


Fig. 2. An example of (a) a 1-IP ($\overline{w} = 9.2$) and (b) a 10-IP ($\overline{w} = 19.8$) for a given pair of reactions ($R00315$ to $R04737$) in *E. coli*. It is easy to notice here that paths in the 10-IP are mainly variants of the 1-IP that share many common vertices and arcs.

this subnetwork corresponds to a collection of reaction chains assimilated to one or several (for $k > 1$) alternative ways to join two given reactions. Fig. 2(b) shows an example of 10-IP. We group all k-IPs (for a given k) from $\mathcal{G}_{int}$, over all possible pairs of source and destination reactions, into the set k-$\mathcal{IP}$.

However, exactly computing the k paths with the smallest neighboring coefficients, given k (at least 1) and two reactions, is a hard computational problem (never mentioned in the literature; see [15] for the closest related problem, whose difficulty is confirmed). We therefore approximately compute the set k-$\mathcal{IP}$, given k, using a heuristic that we obtained by slightly modifying Yen's algorithm [16] to compute the k minimum weighted circuit-free paths in a weighted directed graph. This is an incremental version of the well-known Dijkstra's algorithm [17], and has running time of $O(kn(m + n \times \log n))$, where n is the number of vertices and m is the number of arcs in the network.

2.4 Interest of integrated pathways

As introduced before, the k-IP of two reactions, given k, represents a pool of reaction chains in the metabolic network, whose weights measure the relative proximity of the genes involved (*via* their enzymes) in each chain. In our approach, the k-IPs are computed for all pairs of source and destination reactions, so as to allow a wide analysis of the reaction chains in presence of the associated genomic information. Thus, the set of all IPs (even limited to a fixed k) is not meant to identify, as a whole, some precise collection of biologically meaningful entities, but to enable us to focus on a particular subset of interesting IPs

according to a given (genomic or metabolic) context. The context uses here is about operons. An operon is the set of genes that belong to a basic transcription unit. By nature, its genes are contiguous along a given bacterial genome and share a common biological function [18]. From RegulonDB [19], we selected the operons of *E. coli* that are composed of at least two genes (so that the proximity notion has some sense) and participate in $\mathcal{G}_{int}$. The resulting benchmark represent 16.2% of the operons in *E. coli*.

We extract then the set of operonic reaction chains (or *ORCs*). An *ORC* is a collection of reaction chains from the metabolic network with common initial and final reactions, and whose reactions exactly match the genes of an operon : each gene of the operon produces an enzyme catalyzing at least one reaction in the ORC, and each reaction in that ORC is catalyzed by at least one enzyme produced by a gene in the operon. For short, we say that the ORC *exactly matches* the operon, and we intuitively extend this definition from ORCs to IPs.

We are also interested in the set of multi-Operonic Reaction Chains (or *mORCs*), an extension of ORC, where a collection of reaction chains from the metabolic network with common initial and final reactions, and whose reactions exactly match the genes of two or more operons. For short, we say that the mORC *exactly matches* the collection of operons and extend this definition from mORCs to IPs.

These matching study allows us to quantify the observations according to which genes in operons tend to produce enzymes for consecutive reactions, to study operons in the context of a directed metabolic network (thus identifying the most interesting operons in the sets obtained by preceding studies by [4,2,20] which use undirected metabolic networks), to check the colinearity hypothesis recently investigated by [9].

3 Results

3.1 A description of integrated pathways

The application of our approach on *E. coli* produces 439 382 IPs for each k. As a concrete application, Fig. 2 shows the 1-IP and the 10-IP for the reactions R00315 and R04737. The information about the set k-$\mathcal{IP}$, for $k = 1$ and $k = 10$, is summarized in Tab. 1, together with the corresponding information when the genome is shuffled (notation k-$\mathcal{IP}_{\widetilde{\mathcal{G}}}$), respectively when the metabolism is shuffled (notation k-$\mathcal{IP}_{\widetilde{\mathcal{M}}}$). When less than 10 paths exist to join two given vertices, then the 10-IP contains only the existing paths.

Tab. 1. Description of the sets of integrated pathways: in the integrated network of *E. coli* (first row), its variant obtained by randomly shuffling the genome (second row; average over 10 experiments), and its variant obtained by shuffling the metabolic network (third row; average over 10 experiments). Columns # k-IPs, # genes, # reactions respectively contain the total number of k-IPs in the set k-$\mathcal{IP}$, the average number of genes involved in a k-IP, and the average number of reactions involved in a k-IP. In small characters is shown the variance of each parameter.

Data Set	k	# k-IPs	# genes	# reactions
k-$\mathcal{IP}$	1	439382	11.8 ± 4.5	13.9 ± 5.6
	10		13.8 ± 4.7	17.6 ± 6.2
k-$\mathcal{IP}_{\widetilde{\mathcal{G}}}$	1	439382	11.6 ± 4.6	13.8 ± 5.6
	10		13.6 ± 4.8	18.4 ± 6.8
k-$\mathcal{IP}_{\widetilde{\mathcal{M}}}$	1	651440	7.3 ± 2.2	7.6 ± 2.3
	10		16.6 ± 4.3	14.8 ± 5.1

Column # k-IPs shows that, whereas the number of vertices is the same in the three types of integrated networks, there is a 50% increase of the number of pairs connected by a path in $\mathcal{G}_{int}$ when the metabolic network is shuffled (see k-$\mathcal{IP}_{\widetilde{\mathcal{M}}}$ vs. k-$\mathcal{IP}$); which is justified by the different number and size of strong connected components in the two integrated networks. Moreover, columns # genes and # reactions show that in the integrated network of *E. coli*, the 1-IP is generally much longer than in the integrated networks with shuffled metabolic network (11.8 genes against 7.3), but the other paths in the 10-IP generally add much less vertices to the 1-IP. In other words, the 10 paths with smallest neighboring coefficient in $\mathcal{G}_{int}$ for *E. coli* share a lot of vertices and arcs (see Fig. 2), whereas in the case of shuffled networks those paths are much more different. As a consequence, the 10 paths in $\mathcal{G}_{int}$ for *E. coli* have almost the same length, which explains - together with the fact that only few pairs of vertices are joined by more than 10 paths - why the length of the IPs obtained for *E. coli* and for the examples resulted by genome shuffling are almost the same.

Projecting now this information (obtained, we emphazise it, in the *integrated network* $\mathcal{G}_{int}$) on the metabolic network, we stand that the alternative paths between two given reactions in a metabolic network are mainly closely related variants of a unique path, thus resulting in a structure that we call a *supertrack*. This property is obviously a particularity of the metabolic network, as shown by the results on the shuffled metabolism experiments.

3.2 Operonic insights

For each k from 1 to 10, we compared each operon with the set of genes involved in each k-IP and we computed both the mutual coverage rate according

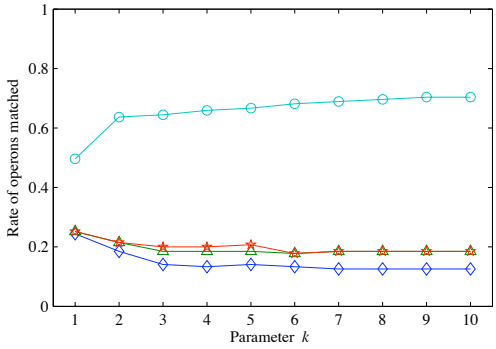


Fig. 3. Operonic interest of k-$\mathcal{IP}$ for distinct k in *E. Coli*. The $\circ$-line represents the rate of operons that are covered by at least one k-IP (coverage rate of the operon by a k-IP equals 1). The $\diamond$-line (the $\triangle$-line and the $\star$-line respectively) resumes the rate of operons that induce ORCs (mORCs with at most 2 and at most 3 operons respectively).

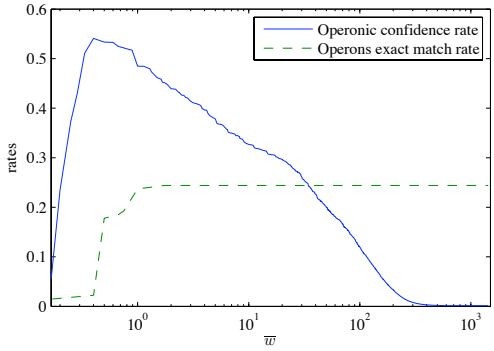


Fig. 4. Rate of 1-IPs that exactly match one operon (thus corresponding to ORCs), with respect to the neighboring coefficient $\overline{w}$ in *E. coli*. Plain line shows the evolution of this rate when $\overline{w}$ is upper bounded by a given value (marked on the X-axis). Complementary, the dashed line represents the rate of operons exactly matching a 1-IP with $\overline{w}$ upper bounded by a given value.

to Jaccard's measure and the coverage rate of the operon by the k-IP. We associated with each operon its best matching k-IP according to Jaccard's measure, and its best matching k-IP according to the coverage rate. Conversely, we associated with each k-IP its best matching operon according to Jaccard's measure.

Relatively many operons correspond to (m)ORCs

With a Jaccard's measure equal to 1, we found that 24.4% of the operons (namely, 33 over 135 operons) match exactly one 1-IP each (i.e. this 1-IP is an ORC, see the $\diamond$-plot in Fig. 3 for $k = 1$), meaning that each such operon produces, using all its genes, all the enzymes needed to catalyze the corresponding reaction chain. With a coverage rate equal to 1, we found also that 49.63% of the operons are completely covered by at least one 1-IP ($\circ$-plot in Fig. 3, with $k = 1$). These rates drop to 12.59 % and rise to 64.44 % respectively when $k = 10$. As k becomes larger, the k-IPs become larger, and thus tend to have associated sets of genes that strictly contain operons ($\circ$-line in Fig. 3 shows increasing values) rather than exactly matching an operon ($\diamond$-line in Fig. 3 shows decreasing values). It is worth noticing here that 8.15% of the operons (11 of them) exactly match a k-IP for *all* values of k from 1 to 10, mainly due to the fact that in this case no alternative path exists to the 1-IP, and thus the 10-IP is identical to the 1-IP.

We then repeated the comparison based on the Jaccard's measure for couples (and respectively triples) of operons and, again, k-IPs. With a Jaccard's measure equal to 1, we found that 14 couples (2 triples, respectively) of operons match exactly one k-IP each, for various $k \geq 1$ (i.e. the projection of this k-IP into the metabolic network is an mORC). In Fig. 3, the $\triangle$-line and the $\star$-line show the corresponding increase in

Tab. 2. Summary of Jaccard's measure between k-IPs ($k = 1$ and $k = 10$) with relatively small $\overline{w}$ and the operons from *E. coli*.

Data Set	k	Rate of operons exactly matched by k-IPs		
		$\overline{w} \leq 1.0$	$\overline{w} \leq 5.0$	$\overline{w} \leq 200.0$
k-$\mathcal{IP}$	1	23.71%	24.44%	24.44%
	10	8.15%	12.59%	12.59%
k-$\mathcal{IP}_{\widetilde{\mathcal{G}}}$	1	0%	0%	2.22%
	10	0%	0%	0.74%
k-$\mathcal{IP}_{\widetilde{\mathcal{M}}}$	1	1.48%	1.48%	1.48%
	10	0%	0%	0%

the rate of operons, when mORCs (with at most two and at most three operons respectively) are considered additionnaly to ORCs. No result was found for four operons or more.

Relatively many 1-IPs with small $\overline{w}$ correspond to operons and ORCs

Complementary to the research of operons that exactly match reaction chains is the research of reaction chains that exactly match operons. In this purpose, we focus on the set 1-$\mathcal{IP}$, as it presents the highest number of exactly mached operons (according to Fig. 3). Fig. 4 details the evolution of the rate of 1-IPs that exactly match operons (or *operonic confidence rate*) when $\overline{w}$ is upper bounded by the value on the X-axis (plain line). Additionnally, the rate of operons that exactly match a 1-IP is drawn (dashed line). The operonic confidence rate increases between the bounds 0 and 0.5, showing two important facts: first, that 1-IPs that match operons (in other words, ORCs) tend to have pairs of successive reactions that are catalyzed by enzymes produced by the same gene (than contributing 0 to the weight of the 1-IP); second, that the gene order along the genome and according to the 1-IP tends to be the same (thus usually contributing 1 to the

weight of the 1-IP, while in the contrary case the contribution would be much more important). Above 0.5, an increase of $\overline{w}$ leads to an initially slow decrease of the operonic confidence rate. The confidence rate exceeds 50% of operon recognition for a bound between 0.4 and 1. The operon with highest $\overline{w}$ has $\overline{w} = 1.75$.

Very few ORCs appear in random networks

Tab. 2 shows the comparison between the rate of operons exactly matching a k-IP ($k = 1$ and $k = 10$) in *E. coli* and in the shuffled data. Clearly, a random gene order (row k-$\mathcal{IP}_{\widetilde{\mathcal{G}}}$) significantly changes the genomic distances between the genes belonging to the same operon of *E. coli*, leading to a very important increase of $\overline{w}$ for the 1-IPs which exactly match the operon. These 1-IPs (and the corresponding ORCs) still exist (the metabolic network didn't change), but one cannot identify them. A random metabolic network (row k-$\mathcal{IP}_{\widetilde{\mathcal{M}}}$) significantly changes the paths with respect to the network of *E. coli*, and the result is a very small number of operons exactly matching a 1-IP only by chance.

4 Discussion

Comparing our results with operons points out the existence, and often the biological significance, of the *supertracks* within a metabolic network (Sec. 3.1). They are the projections into the metabolic network of the k-IPs found in the integrated network $\mathcal{G}_{int}$ (an example of k-IP was shown in Fig. 2). They also revealed to be projections into the metabolic network of about 25% of the operons (thus being qualified as (m)ORCs, see Sec. 3.2 and Fig. 3).

Previous works [21,2,20] assume that the intra-operonic gene order relies on the role of encoded enzymes in the bacterial metabolism. Although they consider undirected metabolic networks and thus accept neighboring relationships between reactions that are not allowed in our approach, these studies show the great need for evidences to support that assumption. [9] first proposes a systematic study, by considering pairs of (not necessarily successive) genes within the same operon and asking whether their operonic order reflects the functional order of the encoded enzymes, as recorded by their participation to a common biochemical pathway. Such pairs are so-called *colinear*, and they fulfill constraints involving the order of genes and reactions only, and not the immediate succession of the genes along the genome or of the reactions along some reaction chain. The study shows that approximately 60% of the gene pairs in *E. coli* are colinear. Turning back to our approach - whose constraints involve both the order and the succession of genes and reactions - *supertracks* are topologically (and not biochemically) defined, however, they do match or include operons. First, this fact shows the tendency of operonic genes to participate together to the same process in the metabolic network, for 24.4% of operons encode precisely the set of enzymes that are necessary to catalyze a reaction chain, whereas 49.6% of them encode the sets of enzymes strictly included in that necessary to catalyze a reaction chain. Second, this combined genomic and metabolic proximity does not necessarily imply colinearity since the gene order within the genome may be entirely reversed (no colinear pair) or merged (some, not all, pairs are colinear) with respect to the order of reactions along the reaction chain. Indeed operons *kbl.tdh, cyn.TSX, csiD.lhgO.gabDTP, otsBA, dgoRKADT* are reversed, whereas the operon *fadIJ* is found both in right and reversed order, and operons *rhaBAD, glgCAP, araBAD, rfbBDACX* have their gene set merged along the reaction chain. However, some operons show a *strong* (topological) colinearity, since all their genes appear exactly in the same order in the genome and, *via* their encoded enzymes, along the reaction chain, even when the reaction chain is much longer than the operon. For instance, the *fadBA* operon contains 2 genes that encode enzymes catalyzing a six reaction chain (*fadB* and *fadA* respectively encode for the first two reactions and the last four reactions). Some other short operons (2 genes) contribute to catalyse short reaction chains (2 genes) thus being colinear pairs (as considered in [9]), but with the notable property of being made of successive genes corresponding to successive reactions (which is not required by a colinear pair and emphasizes again the interest of our approach).

For instance, our approach provides a novel emphasis of the couples or triples of operons defining (m)ORCs, that we identified using k-IPs. Some of them are already known in a regulatory context, for instance *fadBA, fadIJ* that share the dual common repressor *ArcA, fadR* [22]. Among them, the couple *cysDNC, cysJIH* has been already identified as an über-operon [23]. Other operons have already been associated with a unique metabolism of interest, like *atoDAEB* and *fadIJ* that participate in the fatty acid degradation [24]. ORCs emphasize as well operons that share homologous genes (*ascFB*, *bglGFB*), which gives an insight about a reaction chain that is encoded in distinct locations on the genome, showing an abstraction of robustness as proposed by [25].

5 Conclusion

This paper proposes a general framework that integrates gene proximity information from one or several

data sources into a metabolic network. This integration is obtained using a generic notion of distance between genes, that is projected on the metabolic network *via* the encoded enzymes. The resulting integrated network, called $\mathcal{G}_{int}$, is analyzed by computing the k-shortest paths ($k \geq 1$), or k-IPs, between two given reactions involved in this network, where the optimization uses the generic distance between genes. The collection k-$\mathcal{IP}$ of paths obtained for a given k over all pairs of reactions is then filtered according to an appropriate criteria to extract further information about a given genomic or metabolic context.

Our method allowed us to observe that in metabolic networks, the reaction chains that join two given reactions diverge far less than in general networks. Even when 10 such reaction chains exist between two given reactions, which is not the rule, these chains share a lot of intermediate vertices, making them acting like a unique *supertrack* rather than like 10 different ways to join the two given reactions. Interesting supertracks from the metabolic network were discovered by mapping the k-IPs computed in the integrated network onto biologically relevant entities, like operons. We therefore obtained operon-like supertracks (that we called (m)ORCs), each of which map either an operon, or a small group of operons. Supertracks appear therefore as a precisely defined structure to group reactions chains in the metabolic network that are closely related both by their constitutive elements and by their functional features.

Further applications of our approach should test other biological knowledge like KEGG modules, involve alternative distances between genes, and these are the main lines of our future work.

References

[1] M.Y. Galperin and E.V. Koonin. Who's your neighbour? new computational approaches for functional genomics. *Nat. Biotechnol*, (18):609–613, 2000.

[2] Frédéric Boyer, Anne Morgat, Laurent Labarre, Joël Pothier, and Alain Viari. Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. *Bioinformatics*, 21(23):4209–15, Dec 2005.

[3] R.A. Notebaart, B. Teusnik, R. J. Siezen, and B. Papp. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Computational Biology*, (4(1)):e26, 2008.

[4] H. Ogata, W. Fujibuchi, S. Goto, and M. Kanehisa. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.*, (28):4021–4028, 2000.

[5] E.J.B. Williams and D. J. Bowles. Coexpression of neighboring genes in the genome of *arabidopsis thaliana*. *Genome Res.*, (14):1060–1067, 2004.

[6] Jan Ihmels, Sven Bergmann, and Naama Barkai. Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20(13):1993–2003, 2004.

[7] J. Ihmels, R. Levy, and N. Barkai. Principles of transcriptional control in the metabolic network of *saccharomyces cerevisiae*. *Nat. Biotechnol.*, (22(1)):86–92, 2004.

[8] H. Wei, S. Persson, T. Mehta, V. Srinivasasainagendra, L. Chen, G.P. Page, C. Somerville, and A. Loraine. Transcriptional coordination of the metabolic network in *arabidopsis thaliana*. *Plant Physiology*, (18):762–774, 2006.

[9] Károly Kovács, Laurence D Hurst, and Bal·zs Papp. Stochasticity in protein levels drives colinearity of gene order in metabolic operons of *escherichia coli*. *Plos Biol*, 7(5):e1000115, May 2009.

[10] Eduardo P C Rocha. The organization of the bacterial genome. *Annu Rev Genet*, 42:211–33, Jan 2008.

[11] F R Blattner, G Plunkett, C A Bloch, N T Perna, V Burland, M Riley, J Collado-Vides, J D Glasner, C K Rode, G F Mayhew, J Gregor, N W Davis, H A Kirkpatrick, M A Goeden, D J Rose, B Mau, and Y Shao. The complete genome sequence of *escherichia coli k-12*. *Science*, 277(5331):1453–62, Sep 1997.

[12] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, and Akihiro Nakaya. The kegg databases at genomenet. *Nucleic Acids Res*, 30(1):42–6, Jan 2002.

[13] Didier Croes, Fabian Couche, Shoshana J Wodak, and Jacques van Helden. Metabolic pathfinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Res*, 33(Web Server issue):W326–30, Jul 2005.

[14] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–3, May 2002.

[15] C. E. Yang, L. R. Foulds, and J. L. Scott. A pseudo-polynomial algorithm for detecting minimum weighted length paths in a network. *European Journal of Operational Research*, 57(1):123 – 131, 1992.

[16] Jin Y Yen. Finding the *k* shortest loopless paths in a network. *Management Sci.*, 17:712–716, 1970.

[17] E. W Dijkstra. A note on two problems in connexion with graphs. *Numer. Math.*, 1:269–271, 1959.

[18] F Jacob, D Perrin, C Sanchez, and J Monod. L'opéron : groupe de gènes à expression coordonnée par un opérateur. *C R Hebd Seances Acad Sci*, 250:1727–9, Feb 1960.

[19] Heladia Salgado, Socorro Gama-Castro, Martín Peralta-Gil, Edgar Díaz-Peredo, Fabiola Sánchez-Solano, Alberto Santos-Zavaleta, Irma Martínez-Flores, Verónica Jiménez-Jacinto, César Bonavides-Martínez, Juan Segura-Salazar, Agustino Martínez-Antonio, and Julio Collado-Vides. Regulondb (ver-

sion 5.0): *Escherichia coli K-12* transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Research*, 34(Database issue):D394–7, Jan 2006.

[20] Yu Zheng, Joseph D. Szustakowski, Lance Fortnow, Richard J. Roberts, and Simon Kasif. Computational identification of operons in microbial genomes. *Genome Research*, (12):1221–1230, 2002.

[21] H Ogata, Susumu Goto, K Sato, W Fujibuchi, H Bono, and M Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27(1):29–34, Jan 1999.

[22] Ingrid M Keseler, César Bonavides-Martínez, Julio Collado-Vides, Socorro Gama-Castro, Robert P Gunsalus, D Aaron Johnson, Markus Krummenacker, Laura M Nolan, Suzanne Paley, Ian T Paulsen, Martin Peralta-Gil, Alberto Santos-Zavaleta, Alexander Glennon Shearer, and Peter D Karp. Ecocyc: a comprehensive view of *escherichia coli* biology. *Nucleic Acids Res*, 37(Database issue):D464–70, Jan 2009.

[23] Dongsheng Che, Guojun Li, Fenglou Mao, Hongwei Wu, and Ying Xu. Detecting über-operons in prokaryotic genomes. *Nucleic Acids Research*, 34(8):2418–27, Jan 2006.

[24] L S Jenkins and W D Nunn. Genetic and molecular characterization of the genes involved in short-chain fatty acid degradation in escherichia coli: the ato system. *Journal of Bacteriology*, 169(1):42–52, Jan 1987.

[25] Hiroaki Kitano. Biological robustness. *Nat Rev Genet*, 5(11):826–37, Nov 2004.

Weighted-Lasso for Structured Network Inference from Time Course Data

Camille CHARBONNIER[1], Julien CHIQUET[1] and Christophe AMBROISE[1]

STATISTIQUE ET GENOME, UMR 8071 CNRS, 523, place des Terrasses de l'Agora, 91000 Évry France
{Camille.charbonnier, julien.chiquet, christophe.ambroise}@genopole.cnrs.fr

Keywords Biological networks, Transcriptome, Vector auto-regressive model, Weigthed-Lasso

1 Introduction

This model builds upon two popular tools to infer gene regulatory networks from transcriptomic data, namely Gaussian Graphical Models (GGMs) and ℓ_1 regularization. GGMs describe the graph of conditional dependencies between genes while ℓ_1 regularization deals with both high-dimensional setting and selection of the relevant interactions.

We provided in [1] a method that looks for an internal structure of the network in order to drive and improve the selection of edges. Indeed, gene regulatory networks are known not only to be sparse, but also organized, so as genes belong to different classes of connectivity. It thus seems intuitive to search for regulations preferentially between genes where a prior structure suggests they should be.

In a recent paper [2], we extended this approach to VAR1 modeling in order to be able to handle time-course data, understood as one single campaign of repeated measurements over time. We intend here to present this work taking into account recent improvements, as included in the `R` package `SIMoNe` [4] as from version 1.0-0.

2 Modeling structured networks

Our method belongs to the wide class of weighted-LASSO algorithms, designed to reduce false positive discoveries compared to the classical LASSO.

Let us denote by $(X_t = \{X_t^1, \ldots, X_t^p\})_{t\in\mathbb{N}}$ the $\mathbb{R}^p$-valued stochastic process that represents the discrete-time evolution of the p gene expression levels, written as a row vector. Herein, X_t is assumed to be generated by a first-order vector auto-regressive (VAR1) model

$$X_t = X_{t-1}\mathbf{A} + \mathbf{b} + \varepsilon_t, \qquad t \in \mathbb{N}^*,$$

where the noise ε is Gaussian with zero mean and covariance Σ. With adequate assumptions on Σ, X_t is a first order Markov process and each entry A_{ij} is directly linked to the partial correlation coefficient between X_t^i and X_{t-1}^j. Practically, nonzero entries of $\mathbf{A}$ code for the adjacency matrix of a directed graph describing the conditional dependencies between genes.

Matrix $\mathbf{A}$ is inferred by maximizing the ℓ_1-penalized log-likelihood. The main purpose of [2] is to show the interest of taking into account information about the topology of the network. The directedness of the model enables us to distinguigh regulator genes from regulatees and thereby provide the network with an asymmetric structure $\mathbf{Z}$, either inferred or designed upon biological feedbacks. A structure-based weight-matrix $\mathbf{P}^{\mathbf{Z}}$ can then be used to inflate the overall penalty level λ on less probable edges, that is to say edges leaving from regulatees, or deflate it on more probable edges, leaving from regulators.

To sum up, we solve the following penalized log-likelihood maximization problem:

$$\widehat{\mathbf{A}} = \arg\max_{\mathbf{A}} \mathcal{L}(\mathbf{A}; \mathbf{X}) - \lambda \cdot \|\mathbf{P}^{\mathbf{Z}} \star \mathbf{A}\|_{\ell_1},$$

where $\mathcal{L}$ denotes the log-likelihood, λ tunes the overall sparsity of the network, $\mathbf{P}^{\mathbf{Z}}$ is the weight matrix adapted to the internal structure Z of the network, and operator $\star$ represents term-by-term product.

3 Inference

The inference algorithm runs in three steps:

1. Inference of a family of networks deprived of structure ($\mathbf{P}^{\mathbf{Z}} = 1$) for a well chosen set of λ values. We select one of those according to an information criterion (e.g. AIC, BIC) in order to define an initial adjacency matrix $\widehat{\mathbf{A}}^{\text{init}}$.
2. Inference of the structure $\widehat{\mathbf{Z}}$ on $\widehat{\mathbf{A}}^{\text{init}}$ with help of the `R` package `mixer`.
3. Inference of a new family of networks, this time using information about the structure through the weight matrix $\mathbf{P}^{\widehat{\mathbf{Z}}}$.

As compared with the inference method developed in [2] the last two steps have been refined in the `R` implementation thanks to the use of the `R` package `mixer`. This improvement answers the question of

the choice of weights, which can now be derived from the connectivity matrix inferred by `mixer`: the higher the connectivity probability between two classes, the smaller the penalization for edges between genes of these two classes.

4 Application to *E. coli* SOS Network

We focus on a sub-network from *E. Coli* S.O.S. DNA repair network analyzed by [7]. Data provide information on the main 8 genes of the S.O.S. network across 50 time points. This dataset has already been investigated under the light of Bayesian networks by [6] and is well documented. According to the regularly updated EcoCyc database, *lexA* is the only regulator in this sub-network, regulating all genes including itself. We therefore know which network to expect. We compare in Fig. 1 the performances in terms of Precision and Recall rates of the Lasso, the Adaptive Lasso [9], a Bayesian Network based method called G1DBN [5], a recursive elastic-net method called Renet [8], a weighted-Lasso KnownCl which knows *lexA* to be the only hub in the network and finally a weighted-Lasso InferCl with inference of the hub structure $\mathbf{Z}$. Lasso and KnowCl networks are presented in Fig. 2.

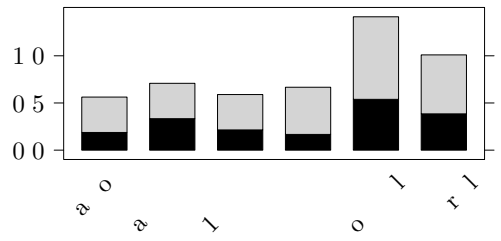
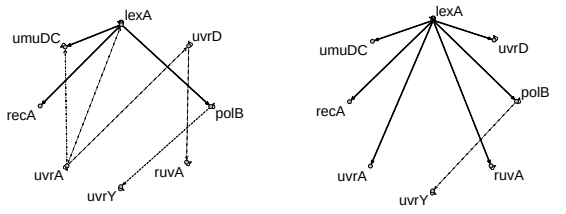


Fig. 1. Precision (black) and Recall (grey) values for different methods on the second experiment of *E. coli* SOS network data from [7].



(a) Lasso (b) Known Hub Clustering

Fig. 2. Graphs inferred with (a) classical Lasso and weighted Lasso with (b) known clustering on the second experiment of *E. coli* SOS network data from [7]. True discoveries are drawn in full lines, False discoveries in dashed lines. Penalty level was chosen according to BIC criterion.

5 Discussion

We propose a weighted-Lasso algorithm designed to tackle time varying gene expression data for which models assuming i.i.d. data become irrelevant. The proposed approach taking into account an underlying structure outperforms similar methods. Even when regulators and regulatees cannot *a priori* been distinguished through analysis of the literature, inference of the classification improves the performances of the Lasso in terms of both recall and precision. It therefore seems good to advice that, whenever available, knowledge about potential transcription factors should be taken into account and that basic knowledge on the topology of biological networks should not be omitted in the modeling process.

Finally, we would like to emphasize the fact that this method is now adapted to the multitask setting (time-course adaptation of [3]) in the new version of the `R` package `SIMoNe`. It will therefore be able to handle dataset pooling time-course observations from different measurement campaigns.

References

[1] C. Ambroise, J. Chiquet, and C. Matias. Inferring sparse Gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238, 2009.

[2] C. Charbonnier, J. Chiquet, and C. Ambroise. Weighted-lasso for structured network inference from time course data. *SAGMB*, 9(1), 2010.

[3] J. Chiquet, Y. Grandvalet, and C. Ambroise. Inferring multiple graphical structures. *Statistics and Computing*, to appear.

[4] J. Chiquet, A. Smith, G. Grasseau, C. Matias, and C. Ambroise. Simone: Statistical inference for modular networks. *Bioinformatics*, 25(3):417–418, 2009.

[5] S. Lèbre. Inferring dynamic genetic networks with low order independencies. *SAGMB*, 8(1), 2009.

[6] B. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alché-Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19, 2003.

[7] M. Ronen, R. Rosenberg, B. Shraiman, and U. Alon. Assigning numbers to the arrows: parametrizing a gene regulation network by using accurate expression kinetics. *PNAS*, 99(16):10555–10560, 2002.

[8] T. Shimamura, S. Imoto, R. Yamaguchi, A. Fujita, M. Nagasaki, and S. Miyano. Recursive regularization for inferring gene networks from time-course gene expression profiles. *BMC Systems Biology*, 3(41), 2009.

[9] S. Zhou, S. van de Geer, and P. Bühlmann. Adaptive lasso for high dimensional regression and Gaussian graphical modeling. *ArXiv*, 2009.

Replication-associated mutational strand asymmetry in the human genome

Chun-Long Chen[1], Benjamin Audit[2], Lauranne Duquenne[1,4], Guillaume Guilbaud[3], Aurélien Rappailles[3], Yves d'Aubenton-Carafa[1], Olivier Hyrien[3], Alain Arneodo[2], & Claude Thermes[1]

[1]CENTRE DE GENETIQUE MOLECULAIRE, FRE115 CNRS, 91198 Gif-sur-Yvette, France
chen@cgm.cnrs-gif.fr, daubenton@cgm.cnrs-gif.fr, thermes@cgm.cnrs-gif.fr
[2]LABORATOIRE JOLIOT CURIE ET LABORATOIRE DE PHYSIQUE, ECOLE NORMALE SUPERIEURE DE LYON, CNRS, 69364 Lyon, France
baudit@ens-lyon.fr, alain.arneodo@ens-lyon.fr
[3]ECOLE NORMALE SUPERIEURE DE PARIS, UMR CNRS 8541, 46 rue d'Ulm 75005 Paris, France
rappaill@biologie.ens.fr, guilbaud@biologie.ens.fr, hyrien@biologie.ens.fr
[4]UMR CNRS 5558, LBBE, UCB Lyon1, 43 bd du 11 Novembre 1918, 69622 Villeurbanne, France
duquenne@biomserv.univ-lyon1.fr

Keywords human genome, nucleotide substitutions, replication, strand asymmetry.

1 Introduction

During evolution, mutations do not occur at the same rate on the two DNA strands. In prokaryotes, they can occur at different rates on the leading and lagging replicating strands and on the transcribed and non-transcribed strands due to asymmetries intrinsic to the replication and transcription processes. In bacterial genomes, unequal intra-strand frequencies of complementary nucleotides have been associated with replication, the leading strand presenting an excess of G over C and/or of T over A [1]. These compositional asymmetries (GC and TA skews defined as S_{GC}=(G-C)/(G+C), S_{TA}=(T-A)/T+A) and S= S_{GC}+S_{TA}) have been associated with different nucleotide substitution rates in the leading and lagging strands [1]. This implies that, on the same strand, complementary substitution rates differ from each other and that these asymmetries switch direction when crossing a replication origin, producing a sharp upward jump of the S profile. In human, studies have shown that replication time is a main determinant of mutation rates [2] but no mutational strand asymmetry has been associated with replication [3]. Here, we demonstrate for the first time the existence of mutational strand asymmetries associated with replication.

2 Results and Discussion

Recent studies have revealed a number (1564) of upward jumps of the S profile (S-jump) that were suggested to coincide with replication initiation zones active in germline cells [4, 5]. We propose that replication induces mutational strand asymmetries that have generated these S-jumps during successive germline divisions. To test this hypothesis, we took advantage of recently determined replication timing profiles of several human cell lines including embryonic stem cells [2, 6, 7]. These profiles present peaks pointing to early replication origins active in the corresponding cell type. Using a multi-scale methodology, we detected the peaks (t-peaks) in each replication timing profile (M&M). The distributions of the distance d separating each t-peak from the closest S-jump showed that for each cell type, numerous peaks are significantly ($P < 0.001$) associated with S-jumps. 863 S-jumps are associated (d<100kbp) with a t-peak of at least one cell type. We hypothesize that the initiation zones associated with these t-peaks are active not only in the corresponding cell type but also in germline cells. We examined the substitution pattern on each side of the S-jumps associated with these peaks (M&M). In intergenic regions most complementary substitution rates differ significantly from each other. This pattern is inverted when shifting from one side of the S-jump to the other. This provides evidence of opposed mutational asymmetries in intergenic regions on both sides of these initiation zones. We examined whether these substitution rates have generated the S-jumps. We computed the skew at equilibrium, S^*, that would be produced after long evolutionary times. The S^* profile presents upward jumps similar to the observed jumps showing that the skew results from the observed substitution rates. However, the mean stationary value is significantly larger than the mean current values strongly suggesting that equilibrium has not been attained.

What mechanism has generated these mutational strand asymmetries? Recent studies showed that most human DNA is transcribed [8]; in particular several types of non-coding Pol II transcripts have been detected in intergenic regions [8]. The intergenic skew could thus result from such Pol II transcription. Following this hypothesis, we show that, in addition to the protein coding transcripts, each side of the upward jumps should be also transcribed in a major mode (R+) divergent from the peaks and a minor mode (R-) converging toward the peaks. However, we demonstrate that the

superimposition of mutational asymmetries associated with these R+ and R- transcripts is not compatible with the observed patterns of substitution rates. In particular, when analyzing the C→T and G→A transition rates, we observe no difference between these rates in R+ introns. This implies that transcription cannot generate any significant difference between these rates, in any region and whatever the corresponding transcription levels. However, these rates are significantly different from each other in R- introns as well as in intergenic regions. This allows us to reject the possibility that the substitution rates observed around the upward jumps result from Pol II transcription only.

Another hypothesis is that in intergenic regions, the observed mutational asymmetries result from replication. On both sides of the corresponding initiation zones, the replication forks are mainly oriented divergently from the *S*-jump centre. Along this hypothesis, this would generate different complementary substitutions rates and would produce exactly the skew observed in intergenic regions. In introns, this skew superimposes to the skew associated with transcription [9]. In full agreement with this hypothesis, we observe, downstream of the upward jumps, an increase of *S* in introns transcribed divergently from the peak centre, i.e. in the same direction as the replication fork progression. In introns transcribed in the opposed direction, this replication-associated skew superimposes to the negative skew associated with transcription. As expected, we observe opposed *S* values in the corresponding regions upstream of the upward jumps finally establishing that all data are in agreement with this hypothesis.

In conclusion, we demonstrate for the first time the existence of replication-associated mutational asymmetries and show that replication is a major driving force that shapes human genome composition.

3 Materials and methods

The replication timing data of Hela cell [2] as well as other cell lines [6, 7] were processed and S50 values were computed as described in [2]. The replication initiation zones (*t*-peak) were detected in each replication timing profile using a continuous wavelet transform. The *S*-jumps detected within the skew profile were retrieved from [4]. The distance *d* between each *S*-jump and the closest *t*-peak was computed for each timing profile. 1000 random simulations were performed to evaluate the significance of the distribution of *d* and the corresponding P-values. Substitutions were tabulated in the human lineage since its divergence with chimpanzee using the macaque and orangutan as outgroups as described in [2]. The method used to compute the nucleotide composition at equilibrium is based on the model of sequence evolution with neighbor-dependent mutations introduced by Arndt et al [10].

References

[1] J.R. Lobry and N. Sueoka, Asymmetric directional mutation pressures in bacteria. Genome Biol., 3: RESEARCH0058, 2002.

[2] C.L. Chen, A. Rappailles, L. Duquenne, M. Huvet, G. Guilbaud, L. Farinelli, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, O. Hyrien*, et al.*, Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res., 20: 447-457, 2010.

[3] M.P. Francino and H. Ochman, Strand symmetry around the beta-globin origin of replication in primates. Mol. Biol. Evol., 17: 416-422, 2000.

[4] M. Touchon, S. Nicolay, B. Audit, E.B. Brodie of Brodie, Y. d'Aubenton-Carafa, A. Arneodo and C. Thermes, Replication-associated strand asymmetries in mammalian genomes: Toward detection of replication origins. Proc. Natl. Acad. Sci. USA, 102: 9836-9841, 2005.

[5] B. Audit, S. Nicolay, M. Huvet, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes and A. Arneodo, DNA replication timing data corroborate in silico human replication origin predictions. Phys. Rev. Lett., 99: 248102, 2007.

[6] R.S. Hansen, S. Thomas, R. Sandstrom, T.K. Canfield, R.E. Thurman, M. Weaver, M.O. Dorschner, S.M. Gartler and J.A. Stamatoyannopoulos, Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc. Natl. Acad. Sci. U S A, 107: 139-144,

[7] R. Desprat, D. Thierry-Mieg, N. Lailler, J. Lajugie, C. Schildkraut, J. Thierry-Mieg and E.E. Bouhassira, Predictable dynamic program of timing of DNA replication in human cells. Genome Res., 19: 2288-2299, 2009.

[8] S. Buratowski, Transcription. Gene expression--where to start? Science (New York, N.Y, 322: 1804-1805, 2008.

[9] P. Green, B. Ewing, W. Miller, P.J. Thomas and E.D. Green, Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet., 33: 514-517, 2003.

[10] P.F. Arndt, C.B. Burge and T. Hwa, DNA sequence evolution with neighbor-dependent mutation. J Comput Biol, 10: 313-322, 2003.

Genetic dissection of post-transcriptional regulation of gene expression

Mathieu CLÉMENT-ZIZA[1], and Andreas BEYER[1]

[1] Biotecnology Center , TU-Dresden, Tatzberg 47-49, D-01307, Dresden, Germany
{mathieu.clement-ziza, andreas.beyer}@biotec.tu-dresden.de

Abstract *Expression QTL studies have been carried out using either transcript or protein abundance to monitor gene expression. However, different biological processes underlie those traits since protein levels are affected by post-transcriptional regulation. In this work, we dissected gene expression traits from which we isolated the post-transcriptional component. We modeled post-transcriptional variation as the residuals after regressing on RNA levels. We integrated published data obtained from a yeast population phenotyped at the transcriptomic and proteomic levels of 137 genes. Mapping this inferred post-transcriptional contribution revealed 36 loci that post-transcriptionaly affected 64 proteins. We identified regulatory hotspots that control many genes, and a candidate master regulator of amino-acid metabolism genes. Our work presents an example of how to disentangle related (yet different) complex traits in order to reveal their genetic basis.*

Keywords: QTL, post-transcription, gene expression, yeast.

1 Introduction

Gene expression is a continuous trait that displays a complex genetic inheritance that involves multiple loci. The technique of expression quantitative trait loci (expression QTL) has helped understanding the genetic basis of these variations. It is a variant of QTL which considers gene expression in a population of genetically diverse individuals as a quantitative trait. Expression QTL studies using either transcript or protein abundance to monitor gene expression have been carried out [1,2]. However, different biological processes control those complex traits: at a steady state, the transcript abundance is mainly dependent on transcription and RNA degradation, whereas protein level is also under the control of post-transcriptional processes, e.g. translation or protein degradation. In this work, we dissected gene expression traits from which we isolated the post-transcriptional component in order to better understand its specific regulation. A QTL analysis of this inferred trait was then carried out to unravel the genetic basis of post-transcriptional regulation of gene expression.

2 Results

The group of Leonid Kruglyak has generated a genetically diverse *Saccharomyces cerevisiae* population derived form a cross between a wild isolate (RM11-1a, hereafter RM) and a laboratory strain (BY4716, hereafter BY). This cross has been used to explore the genetic regulation of either transcript [1] or protein [2] variation. The original publication of the proteomics QTL data just assessed protein concentration as a quantitative trait. However, such analysis ignores the fact that protein concentration variation may just reflect variation of mRNA levels. Hence, it is a priory unknown if the trait reflects variation of transcript levels or post-transcriptional regulation. We sought to separate transcriptional and post-transcriptional regulation.

After combining these two data sets we obtained phenotype data (both transcriptomic and proteomic data) for 137 genes, and genotype data for 1106 informative markers in 93 segregant strains. For each gene, we regressed the proteomic measurements against the transcriptomic ones and considered the residuals as the post-transcriptional contribution to gene expression. We determined linkage between residuals in the segregants and the genetic markers using a novel mapping method developed in the group (manuscript in preparation). This gives rise to what we call post-transcriptional QTL or ptQTL. It allowed us to map 36 loci that contribute to the post-transcriptional regulation of 64 genes (Figure 1).

2.1 Detection of post-transcriptional regulatory hotspots

In the previous expression QTL studies it has been shown that loci that affect gene expression are not evenly distributed throughout the genome and that few hotspot loci can regulate the expression of numerous genes. We detected 2 post-transcriptional regulatory hotspots ($p<0.005$, as described

previously; Brem et al., 2002). One locus is located on chromosome III and mapped to *LEU2*, a gene essential for leucine biosynthesis that had been artificially deleted in the RM parental strain. The other hotspot located on chromosome XIII is also involved in the control of protein levels [3], but the true causative genes had not been identified yet.

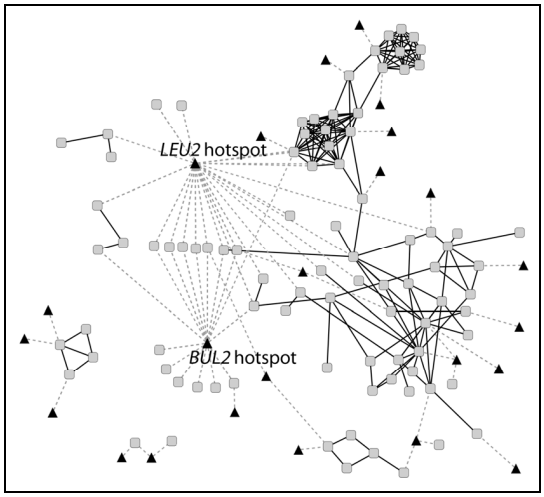


Fig. 1. Protein-protein interaction network of the target genes used in the post transcriptional QTL analysis. Square gray nodes represent a gene/protein. Triangle black nodes correspond to post transcriptional QTL loci. Solid lines represent protein-protein interactions; dashed lines represent QTL linkage.

2.2 A missense polymorphism in BUL2 could underlie the QTL hotspot

Genes affected by the hotspot on chromosome XIII are enriched for genes involved in amino acid (AA) metabolism ($p<0.005$). This hotspot contains the candidate gene *BUL2* which encodes an adaptator component of the Rsp5p ubiquitin ligase complex that regulates the expression, localization, and activity of the high capacity general amino acid permease (Gap1p). Gap1p has been shown to be active at the plasma membrane when internal amino acid levels are low. When internal amino acid concentrations are sufficient, Gap1p is polyubiquinated by the Rsp5p-Bul1p-Bul2p complex and sorted to the vacuole [3]. The amino acid sequence of Bul2p in BY and RM differs by 2 substitutions, one of which affects a highly conserved residue. We hypothesize that Bul2p function could be altered in BY bul2 strains leading to a modified regulation of Gap1p and subsequently a change in amino acid uptake capacity.

2.3 Epistatic interaction between post-transcriptional QTL hotspots

Our results showed that 11 out of 18 genes controlled by the hotspot on chromosome XIII were also linked to the hotspot at the *LEU2* locus. Therefore, we analyzed the residuals of the genes linked to both hotspots and identified strong epistatic interactions between the *LEU2* and *BUL2* loci. The post-transcriptional regulation of all of the 11 genes linked to the two loci was modulated only in the strains carrying both the deletion of *LEU2* (i.e. the RM allele) and the BY *BUL2* allele. Those findings suggests that the post-transcriptional regulation of those genes is the consequence of a modification of amino acid metabolism triggered by both i) leucine limitation due to the *LEU2* deletion, and ii) a modified amino acid uptake capacity due to the altered regulation of Gap1p by BY-Bul2p.

3 Conclusion and perspectives

In this work we demonstrated that the post-transcriptional contribution to gene expression is a quantitative trait affected by multiple natural genetic variations. In order to confirm our predictions we are currently conducting allele switching experiments. We are engineering BY-*bul2* RM and RM-*bul2* BY strains, in which we will asses the phenotypes of the target genes linked to the *BUL2* hotspot through proteomic measurements, and the trafficking of tagged-Gap1p by fluorescence microscopy. This is the first QTL study systematically separating transcriptional and post-transcriptional regulation as distinct traits and our work presents a general framework for disentangling related (yet different) complex traits in order to reveal their genetic basis.

Acknowledgements

This work is supported by the European Commission FP7 grant PhenOxiGen.

References

[1] R. Brem, and L. Kruglyak, The landscape of genetic complexity across 5,700 gene expression traits in yeast. PNAS. 102: 157-1577 ,2005.

[2] E. Foss, D. Radulovic, S. Shaffer, D. Ruderfer, A. Bedalov, D. Goodlett, and L. Kruglyak. Genetic basis of proteome variation in yeast. Nat Genet., 39:1369-1375, 2007.

[3] A. Risinger, and C. Kaiser, Different ubiquitin signals act at the Golgi and plasma membrane to direct GAP1 trafficking. Mol. Biol. Cell, 19:2962-2972, 2008.

Structural and functional genomics in grapevine through FLAGdb^{++}

Sandra Dèrozier[1], Cécile Guichard[1], Franck Samson[1,2], Jean-Philippe Tamby[1], Véronique Brunaud[1], Vincent Thareau[3], Christophe Caron[2,4], Maria Tchoumakov[1], Roberto Bacilieri[5], Anne-Françoise Adam-Blondon[1] and Sébastien Aubourg[1]

[1] Unité de Recherche en Génomique Végétale (URGV), UMR INRA 1165 - Université d'Evry Val d'Essonne - ERL CNRS 8196, CP 5708, F-91057 Evry Cedex
{derozier, guichard, tamby, brunaud, adam, aubourg}@evry.inra.fr
[2] Unité Mathématique Informatique et Génome (MIG), UR INRA 1077, F-78352 Jouy-en-Josas Cedex
fsamson@jouy.inra.fr
[3] Institut de Biotechnologie des Plantes (IBP), UMR CNRS 8618 - Université Paris-Sud, F-91405 Orsay
vincent.thareau@u-psud.fr
[4] Service Informatique et Génomique, Station Biologique, CNRS – UPMC, F-29682 Roscoff Cedex
christophe.caron@sb-roscoff.fr
[5] Unité Diversité et Adaptation des Plantes Cultivées (DiA-PC), UMR INRA 1097, F-34060 Montpellier Cedex 1
roberto.bacilieri@supagro.inra.fr

Keywords genome annotation, database, EuGène, orthology, gene families, integration

Introduction

Three years ago, the whole-genome shotgun sequencing of *Vitis vinifera* opened the way to genomics approaches in grapevine. Indeed, in the framework of a French-Italian consortium, the 487 Mb of a highly homozygous genotype (PN40024) have been sequenced and assembled in a first 8x draft [1]. This work has been improved to release a high quality 12x genome assembly this year. As members of the IGGP consortium, we have provided a curated Vitis gene set essential for the learning of the gene prediction tools. The Genoscope carried out the structural annotation of the genome using the GAZE software as combiner. Based on previous relevant results obtained with the EuGène gene finder tool [2] on other plant genomes, we have decided to use it in order to complete the GAZE annotation and provide a robust and complete gene inventory to the Vitis community. We used the FLAGdb^{++} database [3] for the integration of the Vitis 12x genome and the different structural annotations and functional data.

Results

Among the large panel of gene prediction tools, EuGène presents the advantage to perform both the roles of *ab initio* gene finder and combiner by the integration of several sources of evidence such as splicing sites, protein similarities or cognate transcripts. Furthermore, the weighting of the different evidences that are exploited to predict gene structures is set up through algorithms which explore the parameter space and select the best combination according to a reference gene set. Based on previous practices (see http://eugene.toulouse.inra.fr/), we used EuGène to perform the structural annotation of the 12x grapevine genome. Using a curated set of 600 complete PN40024 genes (genomic regions with cognate full-length cDNAs) and more than 4500 experimentally proved splicing sites, we carried out the training of SpliceMachine for the prediction of the splicing sites and EuGène for the detection of coding regions (Interpolated Markov Model). These data were combined by EuGène to BLASTX and GenomeThreader results taking into account similarities and spliced alignments of more than 400 000 ESTs and cDNAs. Genomic sequences have been masked for repeat elements before gene annotation. Training and annotation tasks have been run on the MIGALE computer platform (MIG unit, Jouy-en-Josas) and final results have been integrated into the FLAGdb^{++} database to complete the Genoscope annotation obtained with GAZE. The evaluation of EuGène results gives sensibility and specificity of 80% and 79% respectively (at the gene scale). Along the 19 Vitis chromosomes, GAZE predicts 26347 genes whereas EuGène predicts 44414 genes. Although this huge difference led us to assume that EuGène over-predicts false positive short genes, we decided to keep all the predicted structures without post-filtering. Actually, out of the 12350 EuGène genes which do not overlap GAZE predictions, 80% of them are suppor-

ted by transcripts and/or similarities on at least half of their size. Furthermore, a similar situation previously obtained in *Arabidopsis thaliana* provides us proofs of the EuGène benefits since the meta-analysis of more than 500 transcriptomes from different organs has experimentally validated 70% of small coding genes that were only detected by EuGène [4].

At the functional level, the standard annotation based on the inference of function by homology has been enriched by the prediction of targeting peptides/signals and membrane domains with a pipeline combining ChloroP, WolfPSORT, Predotar and TMhmm. All the genes have been classified in 2970 families and their phylogenetic profiles have been defined by comparisons against sequence libraries built from 11 phyla.

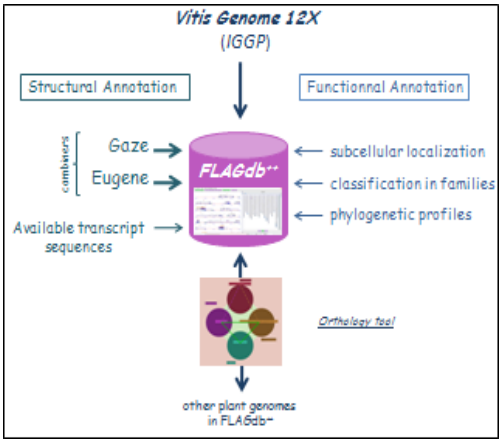


Fig. 1. Annotation and integration around the vitis genome.

Altogether, these predicted and experimental data have been integrated into FLAGdb^{++}. This database, dedicated to plant genomes, is composed of a relational database and an associated user-friendly Java interface [3]. Different tools have been developed with a conceptual effort for the graphical display and the hierarchical organization of the different genomic data in order to browse and explore them and decipher functional relationships between them. We also have written a new Java module allowing the users to access and expertise the predicted links between orthologous genes in the four different plant genomes hosted in FLAGdb^{++}. This tool jointly displays Reciprocal Blast Hit results, intron-exon structures, promoter regions and global protein alignments. Put together, these data permit to reinforce (or invalidate) the orthology relationships, helping therefore the knowledge transfer between Vitis and other plant model species. As a structuring portal for data mining in Vitis, FLAGdb^{++} provides numerous cross-references and links to other databases and tools such as GenBank, URGI and Genoscope genome browsers, PFAM and soon, SNIplay which is dedicated to the sequence diversity in Vitaceae and should be public in the next months.

This level of integration in FLAGdb^{++} has been very useful for the fine characterization of families related to wine characteristics, which have a higher gene copy number (more than twice larger) than in other sequenced plants. Among them, the Stilbene Synthases drive the biosynthesis of resveratrol which has been correlated with the health benefits associated with moderate consumption of red wine (the 'French paradox') and the Terpene Synthases product a high diversity of terpenoids, whereof the relative abundance is correlated with the aromatic traits of wine. Based on GAZE and EuGène annotation, transcript sequences, phylogenies and well known protein features, the 220 members of these families have been detected, annotated and classified. The study of their evolution through comparative genomics approaches and the biochemical characterization of their products are under progress.

Acknowledgements

The authors are grateful to Jérome Gouzy, Thomas Schiex, Philippe Grevet, Aurélie Canaguier, Clémence Bruyère, Joerg Böhlmann and Philippe Hugueney for their support and helpful advices.

References

[1] Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 449(7161):463-7.

[2] Schiex T, Moisan A and Rouzé P (2001) EuGene: An Eucaryotic Gene Finder that combines several sources of evidence. Computational Biology, Eds. O Gascuel and MF Sagot, LNCS 2066, pp. 111-125.

[3] Samson F, Brunaud V, Duchêne S, De Oliveira Y, Caboche M, Lecharny A and Aubourg S (2004) FLAGdb^{++}: a database for the functional analysis of the Arabidopsis genome. Nucleic Acids Research, 32: D347-D350. [http://urgv.evry.inra.fr/FLAGdb]

[4] Aubourg S, Martin-Magniette ML, Brunaud V, Taconnat L, Bitton F, Balzergue S, Jullien PE, Ingouff M, Thareau V, Schiex T, Lecharny A and Renou JP (2007) Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. BMC Genomics, 8:401.

An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications, and transfers

Jean-Philippe DOYON[1], Celine SCORNAVACCA[2], Gergely J. SZÖLLŐSI[3], Vincent Ranwez[4] and Vincent Berry[1]

[1] LIRMM, CNRS - Univ. Montpellier 2, France.
[2] Center for Bioinformatics (ZBIT), Tuebingen Univ., Germany.
[3] LBBE, CNRS - Univ. Lyon 1, France.
[4] ISEM, CNRS - Univ. Montpellier 2, France.

Abstract *(Motivation) Tree reconciliation is an approach that explains the discrepancies between two evolutionary trees by a number of events such as speciations, duplications, transfers and losses. It has important applications in ecology, biogeography and genomics, for instance to decipher relationships between homologous sequences. (Results) We provide a fast and exact reconciliation algorithm according to a parsimony criterion that considers duplication, transfer and loss events. We also present experimental results that give first insights on the conditions under which parsimony is able to accurately infer evolutionary scenarios involving such events. Over all, parsimony performs well under realistic cases, as well as for relatively high duplication and transfer rates. As expected, transfers are in general less accurately recovered than duplications. Availability:* `www.lirmm.fr/phylariane/`

Keywords reconciliation, gene and species trees, transfers, duplications, losses, parsimony.

Un algorithme de parcimonie efficace pour la réconciliation d'arbres de gènes/espèces avec pertes, duplications et transferts

Résumé *(Motivation) La réconciliation d'arbres est une approche qui permet d'expliquer les différences entre deux arbres évolutifs par le biais d'événements comme les spéciations, duplications, transferts et pertes de gènes. Cette approche est appliquée en écologie, en biogéographie et en génomique, par exemple pour étudier les relations entre séquences homologues. (Résultats) Nous proposons un algorithme de réconciliation efficace et exact, basé sur un critère de parcimonie et prenant à la fois en compte les duplications, les transferts et les pertes de gènes. Des résultats expérimentaux montrent que la parcimonie fonctionne bien dans des conditions réalistes, mais aussi dans le cas de taux de duplication et de transfert relativement élevés. Sans surprise, les transferts sont les événements les plus difficiles à inférer correctement.*

Mots-clefs réconciliation, arbres de gènes et d'espèces, transferts, duplications, pertes, parcimonie.

1 Introduction

L'histoire évolutive des organismes vivants est généralement représentée par un *arbre d'espèces* dont les nœuds internes représentent des événements de spéciations [5,21]. L'histoire évolutive d'un ensemble de séquences homologues dérivées d'une séquence ancestrale commune (*famille de gènes*) est elle aussi représentée par un arbre, on parle alors d'un *arbre de gènes*. Contrairement à un arbre d'espèces, un arbre de gènes résulte non seulement d'événement de spéciations, mais aussi de transferts, de duplications et de pertes de matériel génétique. Certains auteurs pensent que les transferts chez les procaryotes (et à proximité de l'ancêtre commun) sont si importants qu'un *réseau de la vie* est plus approprié qu'une simple arborescence [6,7]. Des études complémentaires semblent toutefois indiquer que les transferts n'oblitèrent pas complètement le signal évolutif de spéciation et qu'un arbre de la vie peut encore être discerné malgré le bruit qu'ils engendrent [5,13,21]. Même si ce débat n'est pas encore clos, il a d'ores et déjà engendré des progrès considérables. Par exemple, il est bien établi

que la détection de transferts par approche phylogénétique est plus fiable que par comparaisons de séquences [13,16,26]. L'approche phylogénétique la plus populaire est la *réconciliation d'arbres* et se base sur une comparaison détaillée d'un arbre de gènes avec un arbre d'espèces référent. Ce dernier n'est pas toujours connu mais peut être estimé de manière satisfaisante par des analyses phylogénomiques sur des séquences moléculaires de nombreux gènes ou des caractéristiques de génomes complets [16].

Les méthodes de réconciliation permettent d'expliquer les différences possibles entre un arbre de gènes et un arbre d'espèces suite à des événements de transfert, de duplication et de perte. Une réconciliation d'arbres plonge l'arbre de gènes dans l'arbre d'espèces, représenté par un ensembles de tubes, et associe chaque nœud interne de l'arbre de gènes à un événement évolutif particulier (i.e. spéciation, duplication, transfert ou perte) [19].

Les approches pour réconcilier un arbre de gènes G et un arbre d'espèces S se basent sur des modèles combinatoires [8,19,11,12,9] ou probabilistes [1,25]. Ces derniers intègrent plusieurs paramètres et offrent une meilleure représentation de l'évolution génomique que les modèles combinatoires, mais sont beaucoup plus exigeants en mémoire et temps de calcul. C'est pourquoi seuls les modèles combinatoires sont utilisables pour des études phylogénomiques de plusieurs milliers de familles de gènes [20]. Cependant, les nouvelles technologies permettent d'obtenir les séquences de génomes complets (c.f. [2]) en peu de temps et ces modèles sont en voie de devenir trop lents.

Nous proposons un modèle combinatoire de réconciliation qui considère les événements de spéciation, de duplication, de transfert et de perte (respectivement notés $\mathbb{S}$, $\mathbb{D}$, $\mathbb{T}$, et $\mathbb{L}$), et un algorithme d'une complexité meilleure que ceux actuellement proposés. Formellement, nous considérons le problème d'optimisation nommé *Réconciliation la Plus Parcimonieuse* (ou *MPR* [1]) : pour un arbre d'espèces S, un arbre de gènes G et des coûts associés aux événements $\mathbb{S}$, $\mathbb{D}$, $\mathbb{T}$, et $\mathbb{L}$, trouver une réconciliation de coût minimum (où le coût est la somme des coûts des événements induits par le plongement de G dans S).

Dès que l'on considère les transferts, le problème MPR est NP-complet, même dans le cas où l'on doit réconcilier un seul arbre de gènes binaire avec un arbre d'espèces binaire [12,24]. Ceci est directement lié au fait que les transferts induisent des contraintes chronologiques entre les nœuds de S qui sont difficiles à respecter. En effet, comme les transferts se passent horizontalement entre deux espèces vivant au même moment, ils imposent des contraintes temporelles entre deux nœuds de S (où l'un n'est pas ancêtre de l'autre) qui s'ajoutent aux contraintes initiales (un nœud est nécessairement plus ancien que ses descendants). Ainsi, toute réconciliation ayant plusieurs transferts peut contenir des contraintes temporelles, entre les nœuds de S, mutuellement incompatibles (cf. Fig. 1).

Plusieurs approches ont été proposées pour surmonter la difficulté liée aux contraintes temporelles. Une première solution [10,12] est de définir à l'avance (par des moyens externes à la méthode de réconciliation) les paires de branches de S entre lesquelles les transferts sont autorisés. De nouvelles branches horizontales sont ajoutées pour connecter de telles paires de branches et le graphe obtenu de S est appelé *graphe d'espèces* $\mathcal{S}$. La réconciliation plonge l'arbre de gènes G non plus dans S mais dans $\mathcal{S}$ et une réconciliation la plus parcimonieuse se calcule en temps $O(|\mathcal{S}|^3 \cdot |G|)$. Cependant, calculer un graphe d'espèces induisant une réconciliation la plus parcimonieuse est un problème NP-complet [10]. Une approche plus prometteuse est de considérer une variante réaliste du problème MPR où les nœuds de l'arbre d'espèces sont datés. Ces dates peuvent être calculées par une horloge moléculaire relaxée appliquée sur des arbres de gènes et des séquences moléculaires. Pour le problème de réconciliation, des dates relatives sont suffisantes et la présence de données provenant de fossiles n'est pas requise [15]. Cette piste a été proposée pour des études de coévolution [17,3,18] et est désormais reprise pour la réconciliation d'arbre de gènes/arbre d'espèces [9,24]. La datation des nœuds de S permet

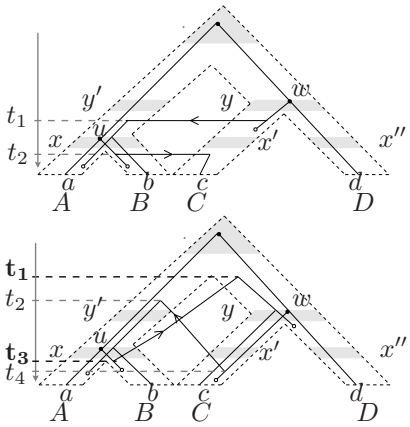


Fig. 1. Deux scénarios de réconciliation entre l'arbre de gènes G (traits pleins) et l'arbre d'espèces S (tubes), où le symbole $\circ$ représente une perte. (En haut) Un scénario temporellement consistant. (En bas) Un scénario temporellement inconsistant : le transfert du donneur au temps t_3 (resp. t_4) au receveur au temps t_1 (resp. t_2) implique que u précède (resp. succède) w.

[1] Cet acronyme est lié à l'intitulé anglophone du problème : "Most Parsimonious Reconciliation".

d'assigner un intervalle de temps à chaque branche. Il est alors possible d'assurer la consistance individuelle de chaque transfert en vérifiant que la branche dite *donneuse* et celle dite *receveuse* ont des intervalles de temps dont l'intersection est non-vide (le transfert est dit temporellement et localement consistant). La variante du problème MPR respectant cette contrainte locale peut être résolue en $O(\max(|S| \cdot |G|)^3)$ par programmation dynamique [18]. Cependant, si deux transferts sont consistants de façon locale mais pas de façon conjointe (cf. Fig. 1), alors la réconciliation n'est pas globablement consistante. De telles inconsistances peuvent être corrigées a posteriori en modifiant certains événements $\mathbb{T}$ [17,18], mais l'optimalité de la réconciliation obtenue n'est plus garantie et l'approche proposée n'est qu'une heuristique pour le MPR.

Une solution pour calculer une réconciliation globalement consistante est de subdiviser la période couverte par S en temps élémentaires, d'associer chacune de ses branches à un de ces temps et de permettre un transfert seulement entre un donneur et un receveur d'un même temps élémentaire. Cette approche permet, dans le cas d'arbres binaires, d'obtenir des algorithmes exacts pour résoudre le problème MPR, comme ceux proposés par [14] et [9]. Le premier a une complexité théorique en $O(|S|^4 \cdot |G|^4)$ tandis que le second est en $O(|S|^4 \cdot k^4 \cdot |G|)$, où k est le nombre de nœuds résultants de la subdivision de S ([2]). Ces complexités, bien que polynomiales, restent élevées et impliquent des temps calcul importants.

Certains des algorithmes décrits ci-dessus s'appuient sur un modèle combinatoire de réconciliation issu de travaux se focalisant sur les duplications et pour lesquels chaque nœud de G est *couplé* avec un seul nœud de S. Toutefois, un tel couplage est insuffisant pour les transferts car il ne peut explicitement indiquer à la fois le donneur et le receveur d'un transfert immédiatement suivi d'une perte. Cette difficulté a conduit certains auteurs à ne considérer qu'une restriction du problème MPR qui néglige le coût des pertes [12,14,24].

Étant donnés un arbre de gènes G et un arbre d'espèces S daté, nous présentons un algorithme polynomial de réconciliation basé sur un modèle combinatoire où les quatre types d'événements évolutifs ($\mathbb{DTLS}$) sont considérés [3]. Contrairement aux approches existantes, notre algorithme gèrent correctement la combinaison d'événements $\mathbb{T} + \mathbb{L}$. Notre modèle s'appuie sur une subdivision S' de S similaire à celle introduite par [9,14,24] et permet de résoudre le MPR en $O(|S'| \cdot |G|)$. Nous explorons ensuite la question fondamentale suivante : *La parcimonie est-elle un critère pertinent pour identifier le véritable scénario évolutif d'une famille de gènes?*

[2] Selon ces auteurs, des modifications non-mentionnées dans le papier permettent d'obtenir une version en $O(|S|^2 \cdot k^2 \cdot |G|)$, mais dont la correction reste à montrer.

[3] L'algorithme considère un coût de spéciation nul, mais il est facile de l'adapter pour un coût non nul.

2 Méthodes

2.1 Définitions et notations basiques

Soit T un arbre où les ensembles de nœuds et de branches sont respectivement notés $V(T)$ et $E(T)$ et seulement ses feuilles sont étiquetées. $r(T)$, $L(T)$ et $\mathcal{L}(T)$ dénotent respectivement sa racine, l'ensemble de ses feuilles et l'ensemble des étiquettes de ses feuilles. Nous allons adopter la convention que la racine est en haut de l'arbre et ses feuilles en bas.

Une branche de T est dénotée $(u, v) \in E(T)$, où u est le père de v. Pour un nœud u de T, T_u dénote le sous-arbre de T enraciné en u, u_p est son père, (u_p, u) est la branche parent de u et $T_{(u_p,u)}$ dénote le sous-arbre de T enraciné avec la branche (u_p, u). Un nœud interne u de T a un ou deux fils, notés respectivement $\{u_1\}$ ou $\{u_1, u_2\}$. Il est important de souligner qu'un arbre T est non-ordonné et les deux fils u_1 et u_2 d'un nœud interne u de T sont interchangeables. Autrement dit, u_1 peut être arbitrairement sélectionné comme l'unique fils de u qui respecte une contrainte donnée. Pour deux nœuds u et u' de T, u' est dit un *descendant* (resp. strict) de u si u est sur l'unique chemin entre u' et $r(T)$ (resp. et $u \neq u'$).

Un nœud interne u de T est dit *artificiel* lorsqu'il a un seul fils. La *contraction* d'un nœud artificiel signifie que ce nœud est enlevé de l'arbre et que les deux branches adjacentes sont jointes. Un arbre T' est dit une *subdivision* d'un arbre T si la contraction récursive de tous les nœuds artificiels de T' donne T.

Un *arbre d'espèces* S est un arbre binaire tel que chaque élément de $\mathcal{L}(S)$ représente une espèce existante et étiquette exactement une feuille de S (il y a une bijection entre $L(S)$ et $\mathcal{L}(S)$). Un *arbre de gènes* G est un arbre binaire. Dorénavant, nous considérons un arbre d'espèces S et un arbre de gènes G tel que $\mathcal{L}(G) \subseteq \mathcal{L}(S)$ et $\mathcal{L} : L(G) \to L(S)$ dénote la fonction qui couple chaque feuille de G à l'unique feuille de S avec la même étiquette. Aussi, le terme arc réfère à une branche de G et le terme branche est pour S.

Dans le reste de l'article, nous supposons que de l'information temporelle est donnée pour l'arbre d'espèces S (c'est-à-dire qu'une période de temps est associée à chaque événement de spéciation) et que l'arbre S est ultramétrique.

Une fonction d'étiquetage temporel pour S est notée $\theta_S : V(S) \to \mathbb{R}$ et est telle que pour chaque feuille

$x \in L(S)$, $\theta_S(x) = 0$, et pour chaque paire de nœuds $x, x' \in V(S)$, que x' soit un descendant strict de x implique $\theta_S(x') < \theta_S(x)$. Cet étiquetage temporel est interprété de la façon suivante : chaque feuille de S correspond à une espèce contemporaine qui existe au temps présent $t = 0$ et chaque nœud interne correspond à une espèce ancestrale qui a donné naissance à deux lignées au temps passé $t > 0$.

DEFINITION 2.1. *Soit un arbre T et un sous-ensemble de feuillles $K \subseteq L(T)$.* L'arbre homéomorphique *de T qui connecte K est noté $T|_K$ et est le plus petit sous-arbre induit de T tel que $L(T|_K) = K$.*

Nous introduisons ci-dessous le concept d'un scénario d'évolution d'un gène débutant à $r(S)$ et évoluant dans S par des événements $\mathbb{DTLS}$. Un tel scénario génère un *arbre de gènes complet* noté G^o, où l'ensemble de feuilles est formé de gènes contemporains mais aussi de gènes perdus durant le scénario (cf. Fig. 1 et 2). Formellement, $L(G^o) = L_{\mathbb{C}}(G^o) \cup L_{\mathbb{L}}(G^o)$, où $L_{\mathbb{C}}(G^o)$ et $L_{\mathbb{L}}(G^o)$ sont disjoints et correspondent respectivement aux gènes contemporains ($\mathbb{C}$) et perdus ($\mathbb{L}$).

DEFINITION 2.2. *Soit un arbre de gènes observé G et un arbre d'espèces S, avec sa fonction d'étiquetage temporel θ_S. Un* scénario $\mathbb{DTLS}$ *pour G le long de S est noté (G^o, M, θ^o_G), où G^o est l'arbre de gènes complet, $M : V(G^o) \to V(S)$ couple chaque nœud u de G^o à un nœud de S et $\theta^o_G : V(G^o) \to [0, \theta_S(r(S))]$ associe chaque noeud u de G^o à une étiquette temporelle de S. Les événements $\mathbb{DTLS}$ correspondants et associés aux nœuds $u \in V(G^o)$ sont définis ci-dessous.*

1. *Si $M(u) = x$, $M(u_1) = x_1$ et $M(u_2) = x_2$, alors u est un événement $\mathbb{S}$.*
2. *Si $M(u) = M(u_1)$ et $M(u) = M(u_2)$, alors u est un événement $\mathbb{D}$.*
3. *Si u est une feuille de G^o qui n'est pas dans G, alors u est un événement $\mathbb{L}$*
4. *Si $M(u_1) = M(u) = x$, $M(u_2) = y$ et y n'est ni un ancêtre, ni un descendant de x, alors u est un événement $\mathbb{T}$, (x_p, x) et (y_p, y) correspondant respectivement aux branches donneuse et receveuse.*

Un scénario $\mathbb{DTLS}$ est dit consistant *si et seulement si les contraintes suivantes sont respectées. Premièrement, l'arbre de gènes homéomorphique $G^o|_{L_{\mathbb{C}}(G^o)}$ est G. Deuxièmement, pour un événement $\mathbb{T}$ tel que décrit ci-dessus (c'est-à-dire en (4)) $[\theta_S(x), \theta_S(x_p)] \cap [\theta_S(y), \theta_S(y_p)] \neq \emptyset$. Troisièmement, pour chaque arc $(u_p, u) \in E(G^o)$, $\theta^o_G(u_p) > \theta^o_G(u)$.*

Le coût d'un tel scénario est noté $Coût(G^o, M, \theta^o_G) = d\delta + t\tau + l\lambda$, où d, t, et l dénotent respectivement le nombre d'événements $\mathbb{D}$, $\mathbb{T}$ et $\mathbb{L}$, et δ, τ et λ sont leurs coûts respectifs.

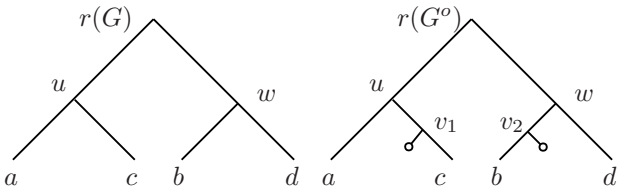


Fig. 2. (a) Un arbre de gènes G avec quatre feuilles a, b, c, et d, appartenant resp. aux espèces contemporaines A, B, C et D (cf. Fig. 1). (b) Un arbre de gènes complet G^o, où $\circ$ représente des gènes feuilles perdus.

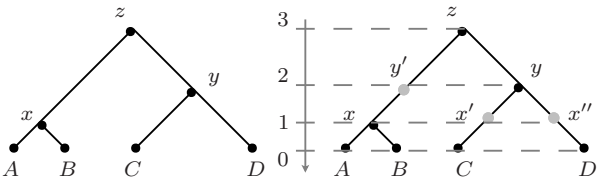


Fig. 3. (a) L'arbre d'espèces S et (b) sa subdivision S'. Les nœuds artificiels de S' sont en gris et dénotés y', x' et x'', où $\theta'_{S'}(x) = \theta'_{S'}(x') = \theta'_{S'}(x'')$ et $\theta'_{S'}(y) = \theta'_{S'}(y')$.

Le problème d'optimisation considéré est nommé *MPR* et est défini ci-dessous.

Entrées. Un arbre d'espèces S avec une fonction d'étiquetage temporel $\theta_S : V(S) \to \mathbb{R}$, un arbre de gènes observé G, la fonction d'association entre feuilles $\mathcal{L} : L(G) \to L(S)$ et les trois coûts δ, τ et λ des événements $\mathbb{DTL}$.

Résultats. Un scénario $\mathbb{DTLS}$ consistant (G^o, M, θ^o_G) pour G le long de S qui minimise $Coût(G^o, M, \theta^o_G)$.

2.2 Un modèle de réconciliation efficace

Pour obtenir un modèle efficace, l'arbre d'espèces est subdivisé pour obtenir une discrétisation du temps (cf. Fig. 3) et permettre de calculer une réconciliation la plus parcimonieuse (de manière similaire à [3,25]).

DEFINITION 2.3. *Pour un arbre (binaire) d'espèces S et une fonction d'étiquetage $\theta_S : V(S) \to \mathbb{R}$, soit S' la subdivision de S suivante : pour chaque nœud $x \in V(S) \setminus L(S)$ et chaque branche$(y_p, y) \in E(S)$ tel que $\theta_S(y_p) > \theta_S(x) > \theta_S(y)$, un nœud artificiel est inséré sur la branche (y_p, y) au temps $\theta_S(x)$. La subdivision nous permet de définir une fonction d'étiquetage temporel pour S' en se basant seulement sur sa topologie : pour chaque $x \in V(S'), \theta'_{S'}(x)$ est le nombre de branches qui séparent x d'une de ses feuilles descendantes (toutes à la même distance).*

L'étiquetage temporel d'une branche (x_p, x) de S' est noté $\theta'_{S'}(x_p, x) = \theta'_{S'}(x)$. Pour un temps t, $E_t(S') = \{(x_p, x) \in E(S') : \theta'_{S'}(x_p, x) = t\}$ dénote l'ensemble des branches de S' localisées au temps t.

Notre modèle de réconciliation défini ci-dessous se base sur six événements et groupes d'événements $\mathbb{DTLS}$, incluant un événement dit "null" et noté $\varnothing$ (cf. Fig. 4).

DEFINITION 2.4. *Une* réconciliation *entre G et S est notée α et associe chaque arc $(u_p, u) \in E(G)$ à une séquence ordonnée de branches de la subdivision S' et notée $\alpha(u_p, u)$. Dans cette séquence de ℓ éléments, $\alpha_i(u_p, u)$ dénote le i-ième élément pour $1 \leq i \leq \ell$. Chaque branche $\alpha_i(u_p, u)$, dénotée ci-dessous (x_p, x), respecte une et seulement une des contraintes suivantes (cf. Fig. 4).*

Premièrement, considérons que (x_p, x) est la dernière branche $\alpha_\ell(u_p, u)$ de la séquence. Si u est une feuille de G, alors x est l'unique feuille de S' ayant la même étiquette que u (c'est-à-dire que $x = \mathcal{L}(u)$) (Contrainte de couplage contemporain). Sinon, un des cas ci-dessous est vérifié.

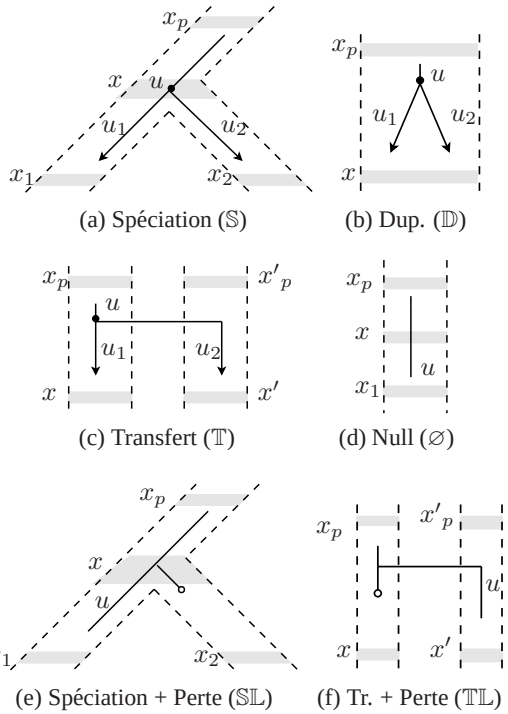
- $\{\alpha_1(u, u_1), \alpha_1(u, u_2)\} = \{(x, x_1), (x, x_2)\}$ *(événement $\mathbb{S}$).*
- *$\alpha_1(u, u_1)$ et $\alpha_1(u, u_2)$ sont tous les deux égales à (x_p, x) (événement $\mathbb{D}$).*
- *$\{\alpha_1(u, u_1), \alpha_1(u, u_2)\} = \{(x_p, x), (x'_p, x')\}$, où (x'_p, x') est une branche de S' différente de (x_p, x) et localisée au temps $\theta'_{S'}(x_p, x)$ (événement $\mathbb{T}$).*

Si (x_p, x) n'est pas la dernière branche $\alpha_\ell(u_p, u)$ de la séquence, un des cas suivants est vérifié.

- *x est un nœud artificiel de S' avec un seul fils x_1, et la prochaine branche $\alpha_{i+1}(u_p, u)$ est (x, x_1) (événement $\varnothing$).*
- *x n'est pas un nœud artificiel et $\alpha_{i+1}(u_p, u) \in \{(x, x_1), (x, x_2)\}$ (événement $\mathbb{SL}$).*
- *$\alpha_{i+1}(u_p, u) = (x'_p, x')$ est différente de (x_p, x) et localisée au temps $\theta'_{S'}(x_p, x)$ (événement $\mathbb{TL}$).*

Une réconciliation α entre G et S' (cf. Fig. 2 et 3) est représentée à la Fig. 1, où le chemin $\alpha(w, b)$ dans S' est $[(y, x'), (y', x), (x, B)]$. Notons que l'arbre de gènes étendu G^o (cf. Fig. 2) est induit de α.

Nous montrons que le modèle (Def. 2.4) permet d'inférer des scénarios optimaux et consistants en temps (Def. 2.2). Premièrement, chaque événement $\mathbb{T}$ se fait entre branches d'une même tranche de temps. Ensuite, chaque perte est couplée avec soit une spéciation ($\mathbb{SL}$) soit un transfert ($\mathbb{TL}$). Considérer une perte seule générerait des réconciliations qui ne seraient pas les plus parcimonieuses : pour un arbre G^o généré par le modèle courant, une seule feuille perdue $l \in L_{\mathbb{L}}(G^o)$ pourrait être remplacée par un sous-arbre sans espèce contemporaine, avec au moins deux feuilles perdues et donc moins parcimonieux. Donc, toute combinaison d'événements $\mathbb{DTLS}$ d'un scénario peut être généré par le modèle, excepté les combinaisons qui ne sont pas parcimonieuses : une spéciation d'un gène où les deux fils n'ont aucun survivant parmi les feuillles de S'; une duplication d'un gène où au moins une des copies s'éteint; un transfert où la lignée de gène transférée s'éteint.



(a) Spéciation ($\mathbb{S}$) (b) Dup. ($\mathbb{D}$)
(c) Transfert ($\mathbb{T}$) (d) Null ($\varnothing$)
(e) Spéciation + Perte ($\mathbb{SL}$) (f) Tr. + Perte ($\mathbb{TL}$)

Fig. 4. Les six événements $\mathbb{DTLS}$ de la Déf. 2.4, où un arc (u_p, u) (ligne pleine) de l'arbre de gènes complet G^o est plongé dans une branche (x_p, x) (tube pointillé; zone blanche) de la séquence $\alpha(u_p, u)$.

Les six cas de notre modèle permettent de progresser soit dans les tranches de temps de S' soit dans les arcs de G : dans toute réconciliation la plus parcimonieuse, un événement de $\mathbb{TL}$ peut être suivi seulement par un des cinq autres événements. Alors, le modèle offre tous les ingrédients pour un algorithme de programmation dynamique pour calculer un scénario consistant en temps et le plus parcimonieux, ce en temps polynomial selon $|S'|$ et $|G|$. Le modèle permet de résoudre le problème MPR de façon efficace et exacte.

2.3 Un algorithme efficace pour MPR

Basé sur notre modèle de réconciliation, nous proposons dans cette section un algorithme polynomial en temps et en espace pour résoudre le problème MPR (cf. Algo. 1 et Thm. 2.5).

Considérons un arc $(u_p, u) \in E(G)$, une branche $(x_p, x) \in E(S')$ et un temps $t = \theta'_{S'}(x_p, x)$.

Notons par $Coût(u, x)$ le coût minimal parmi toutes les réconciliations entre $G_{(u_p,u)}$ et la forêt de sous-arbres de S' enracinés en une branche localisée au temps t, tel que (x_p, x) est la première branche dans la séquence associée à (u_p, u). $Coût(r(G), r(S'))$ correspond au coût minimal d'une réconciliation entre G et S'. L'algorithme de programmation dynamique (voir

le pseudo-code) remplit la matrice $Coût : V(G) \times V(S') \rightarrow \mathbb{N}$ avec deux boucles imbriquées : une qui visite tous les arcs de G selon un parcours de bas-en-haut et une qui visite toutes les étiquettes temporelles de S' en débutant au temps présent $t = 0$ et en remontant progressivement le temps. Pour l'arc (u_p, u) et le temps t actuellement visités (respectivement aux lignes 3 et 4), deux boucles consécutives sur toutes les branches $(x_p, x) \in E_t(S')$ calculent le coût minimal de coupler (u_p, u) avec (x_p, x) selon les six événements $\mathbb{S}$, $\mathbb{D}$, $\mathbb{T}$, $\varnothing$, $\mathbb{SL}$ et $\mathbb{TL}$ (Fig. 4). Pour une branche $(x_p, x) \in E_t(S')$, la première boucle (lignes 5 à 18) calcule le coût minimal pour les cinq premiers événements, la deuxième boucle (lignes 19 à 22) calcule ce coût pour $\mathbb{TL}$ et $Coût(u, x)$ est le coût minimum des six événements.

Algorithm 1 Calcule $Coût(r(G), r(S'))$.

1: Construire la subdivision S' de S de la façon décrite à la Définition 2.3
2: La matrice $Coût : V(G) \times V(S') \rightarrow \mathbb{N}$ est initialisée ci-dessous: si $u \in L(G), x \in L(S')$ et $\mathcal{L}(u) = x$, alors $Coût(u, x) \leftarrow 0$. Sinon, $Coût(u, x) \leftarrow \infty$.
3: **pour tout** $(u_p, u) \in E(G)$ selon un parcours de bas-en-haut **faire**
4: **pour tout** $t \in \{0, 1, \ldots, \theta'_{S'}(r(S'))\}$ **faire**
5: **pour tout** $(x_p, x) \in E_t(S')$ **faire**
6: **si** $u \in L(G), x \in L(S')$ et $\mathcal{L}(u) = x$ **alors**
7: Sauter les lignes 8 à 22 et se rendre à la prochaine itération de la boucle à la ligne 5 {Case de base}
8: $Coût_g \leftarrow \infty$, pour $g \in \{\mathbb{S}, \mathbb{D}, \mathbb{T}, \varnothing, \mathbb{SL}\}$
9: **si** u a deux enfants **alors**
10: **si** x a deux enfants **alors**
11: $Coût_{\mathbb{S}} \leftarrow \min\{Coût(u_1, x_1) + Coût(u_2, x_2), Coût(u_1, x_2) + Coût(u_2, x_1)\}$
12: $Coût_{\mathbb{D}} \leftarrow Coût(u_1, x) + Coût(u_2, x) + \delta$
13: $(y_p, y) \leftarrow MeilleurReceveur((u, u_1), (x_p, x))$
14: $(z_p, z) \leftarrow MeilleurReceveur((u, u_2), (x_p, x))$
15: $Coût_{\mathbb{T}} \leftarrow \min\{ Coût(u_1, x) + Coût(u_2, z), Coût(u_1, y) + Coût(u_2, x) \} + \tau$
si x a un seul enfant **alors** $Coût_{\varnothing} \leftarrow Coût(u, x_1)$
16: **si** x a deux enfants **alors**
17: $Coût_{\mathbb{SL}} \leftarrow \min\{Coût(u, x_1), Coût(u, x_2)\} + \lambda$
18: $Coût(u, x) \leftarrow \min\{Coût_g : g \in \{\mathbb{S}, \mathbb{D}, \mathbb{T}, \varnothing, \mathbb{SL}\}\}$
19: **pour tout** $(x_p, x) \in E_t(S')$ **faire**
20: $(x'_p, x') \leftarrow MeilleurReceveur((u_p, u), (x_p, x))$
21: $Coût_{\mathbb{TL}} \leftarrow Coût(u, x') + \tau + \lambda$
22: $Coût(u, x) \leftarrow \min\{ Coût(u, x), Coût_{\mathbb{TL}} \}$
23: **retourner** $Coût(r(G), r(S'))$

Le cas de la Fig. 4c est considéré aux lignes 13 à 15, où le coût d'un événement $\mathbb{T}$ débutant sur la branche (x_p, x) est calculé pour l'arc (u_p, u). Si (u, u_1) (resp. (u, u_2)) est la lignée de gène transférée, une procédure nommée *MeilleurReceveur* calcule la branche (y_p, y) (resp. (z_p, z)) qui minimise $Cost(u_1, y)$ (resp. $Cost(u_2, z)$) parmi toutes les branches de S' localisées au temps t et différente de (x_p, x). La même procédure est utilisée à la ligne 20 pour le cas $\mathbb{TL}$. Une optimisation similaire pour calculer le meilleur receveur d'un transfert a été trouvée de façon indépendante dans [23].

THEOREM 2.5. *L'Algorithme 1 résout le problème MPR en temps et en espace* $\Theta(|S'| \cdot |G)$.

3 Résultats expérimentaux

Pour un grand nombre d'arbres de gènes simulés, nous avons évalué les performances de notre approche en comparant les scénarios proposés par notre algorithme avec les vrais scénarios évolutifs des arbres. Chaque paire arbre de gènes/vrai scénario est obtenu par un modèle probabiliste d'évolution qui inclue des duplications, des transferts et des pertes.

3.1 Simulation des arbres d'espèces

Nous avons utilisé un processus de naissance et de mort (birth and death) pour générer aléatoirement 10 arbres d'espèces contenant chacun 100 taxa (programme PhyloGen [22] avec un ratio de naissance/mort fixé à 1.25). Ces arbres ont ensuite été normalisés afin qu'ils aient tous la même hauteur h.

3.2 Simulation des scénarios $\mathbb{DTL}$

À partir d'une seule copie d'un gène, présente à la racine d'un arbre S au temps $t = h$, nous avons généré des scénarios $\mathbb{DTL}$ en faisant évoluer cette copie selon un processus de Poisson caractérisé par trois paramètres : le taux de duplication (r_δ), le taux de transfert (r_τ) et le taux de perte (r_λ). Dans le cas d'un transfert, le donneur est choisi uniformément parmi les gènes existants au moment du transfert. On obtient ainsi, pour chaque simulation, un arbre de gènes G^o et une réconciliation simulée α_R incluant les événements $\mathbb{DTLS}$ à l'origine de G^o.

Csűrös et Miklós ont récemment publié une étude portant notamment sur l'ampleur relative des taux de duplication, de transfert et de perte chez les archébactéries [4]. Ils estiment qu'environ 23% des événements sont des duplications, 1% sont des acquisitions, et 76% sont des pertes. Ils observent également un taux approximatif de perte de 1.5 pour un arbre d'hauteur unitaire.

En nous appuyant sur ces résultats, nous avons fait varier de manière réaliste les taux $\mathbb{D}$,$\mathbb{T}$ et $\mathbb{L}$, et créé deux jeux de données. Le premier jeu de données, nommé ds_1, est décrit comme suit : le taux de perte r_λ est 0.7, la hauteur des arbres d'espèces (h) est 1 et les taux r_δ et r_τ varient dans l'intervalle $[0.01, 0.35]$ avec un pas de 0.034 (soit 11 valeurs). Nous avons donc obtenu 11×11 ensembles de paramètres cohérents avec une évolution le long d'une échelle temporelle

importante correspondant, par exemple, au phylum des bactéries ou à celui des archébactéries. En effet, le taux de perte choisi est réaliste (selon [4]) et nous ne faisons pas de suppositions sur le taux relatif de transfert et de perte, la seule contrainte étant que $r_\delta + r_\tau \leq r_\lambda$. Pour chacun des 10 arbres d'espèces et des 121 ensembles de paramètres, nous avons généré 5 arbres de gènes pour un total de 6 050.

Le deuxième jeu de données, nommé ds_2, fixe le rapport $r_\lambda/(r_\lambda + r_\delta + r_\tau)$ à 0.7 [4]. L'objectif ici est d'étudier la pertinence d'une approche de parcimonie pour différentes échelles temporelles (phylogénies profondes ou récentes) en variant la hauteur h de S comme suit : $h = 0.2, 0.4, 0.8$ et 1.6. Le taux de transfert r_τ varie dans l'intervalle $[0, 0.3]$ avec un pas de 0.03 (soit 11 valeurs) et en imposant $r_\delta = 0.3 - r_\tau$. Pour chacune de ces 44 combinaisons de paramètres et les 10 arbres d'espèces, 20 arbres de gènes ont été générés pour un total de 8 800.

3.3 Analyses et résultats

Pour chaque jeu de données, nous avons utilisé comme coût d'un événement $\mathbb{DTL}$ l'inverse du taux moyen de ce type d'événement au long du processus de simulation (par exemple, δ est fixé à $1/0.18$ pour ds_1). Pour chaque couple d'arbres G et S, nous avons calculé une réconciliation parmi les plus parcimonieuses, nommée α_p, grâce à l'Algorithme 1.

Il faut noter qu'une réconciliation réelle α_R contient souvent des événements qui n'ont laissé aucune trace, il est donc impossible à une approche de parcimonie de les retrouver. Par exemple, les duplications immédiatement suivies par une ou plusieurs pertes ou événements $\mathbb{TL}$ consécutifs. Afin de comparer α_p avec le vrai scénario évolutif α_R, nous avons éliminé de celui-ci ces événements dits fantômes et défini une nouvelle réconciliation notée α'_R.

Nous avons d'abord étudié les conditions dans lesquelles la parcimonie peut correctement estimer les événements $\mathbb{DTL}$ en comparant les coûts de α_p et α'_R : quand ils diffèrent de façon importante, la parcimonie n'est plus une approche souhaitable. Le surcoût relatif de α'_R par rapport à une réconciliation la plus parcimonieuse est défini ainsi :

$$Surcoût(\alpha'_R, \alpha_P) = \frac{Coût(\alpha'_R) - Coût(\alpha_P)}{Coût(\alpha_P)}.$$

Il faut noter que $Coût(\alpha'_R) = Coût(\alpha_P)$ n'implique pas $\alpha_P = \alpha'_R$ puisque plusieurs réconciliations plus parcimonieuses peuvent exister. La Fig. 5 montre l'ampleur du surcoût selon les taux r_δ et r_τ et la hauteur de l'arbre. On remarque que le surcoût reste très limité pour toutes les combinaisons de taux mais qu'il augmente sensiblement avec la hauteur de l'arbre. Ceci est probablement dû aux événements cachés de α_R qui sont encore présents dans α'_R.

Nous nous sommes ensuite penchés sur les conditions dans lesquelles la parcimonie retrouve correctement la position des événements $\mathbb{DTL}$ qui ont engendré G. Rappelons qu'une réconciliation α pour un arbre de gènes G définit les événements $\mathbb{DTL}$ associés aux nœuds et branches internes de G. Puisque la position des duplications et des transferts indique univoquement celle des pertes, nous nous sommes focalisés sur les événements $\mathbb{D}$ et $\mathbb{T}$.

Soit $\mathbb{D}_S(\alpha)$ le sous-ensemble de paires $(u, (x_p, x)) \in V(G) \setminus L(G) \times E(S)$ tel que u est une duplication localisée sur (x_p, x) selon α. Soit $\mathbb{T}_S(\alpha)$ le sous-ensemble de triplets $((u_p, u), (x_p, x), (y_p, y)) \in E(G) \times E(S)^2$ tel que (u_p, u) est transféré et (x_p, x) (resp. (y_p, y)) est le donneur (resp. receveur). Pour une réconciliation la plus parcimonieuse α_P, la précision avec laquelle elle retrouve les événements $\mathbb{D}$ et $\mathbb{T}$ de la réconciliation

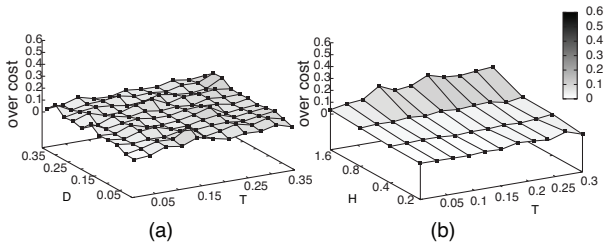


Fig. 5. Le surcoût relatif de α'_R en terme de coût de parcimonie par rapport à une réconciliation parmi les plus parcimonieuses, en faisant varier les taux de duplication et transfert et la hauteur de l'arbre, i.e., (a) ds_1 et (b) ds_2. Les valeurs élevées montrent les cas où il est inadéquat d'utiliser la parcimonie.

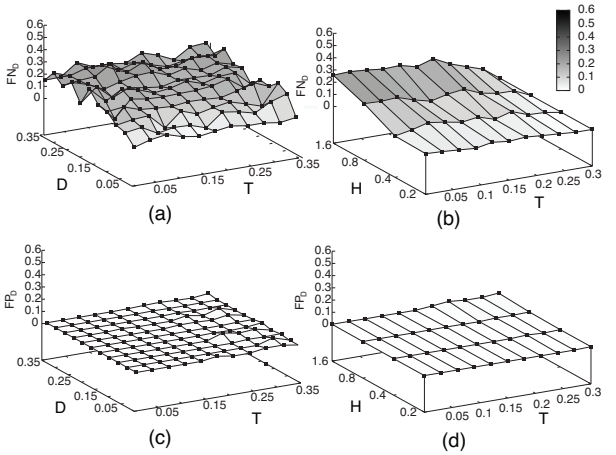


Fig. 6. Étude des conditions dans lesquelles la parcimonie retrouve précisément les événements $\mathbb{DTL}$. Ratios de faux négatifs (a-b) et faux positifs (c-d) pour les événement $\mathbb{D}$ en faisant varier r_δ, r_τ et la hauteur de l'arbre, pour ds_1 (a-c) et ds_2 (b-d).

réelle α'_R est évaluée par les ratios de faux positifs/négatifs définis ci-dessous (où $\mathbb{E} \in \{\mathbb{D}, \mathbb{T}\}$).

$$FP_{\mathbb{E}}(\alpha'_R, \alpha_P) = \frac{|\mathbb{E}_S(\alpha_P) - \mathbb{E}_S(\alpha'_R)|}{|\mathbb{E}_S(\alpha_P)|}$$
$$FN_{\mathbb{E}}(\alpha'_R, \alpha_P) = \frac{|\mathbb{E}_S(\alpha'_R) - \mathbb{E}_S(\alpha_P)|}{|\mathbb{E}_S(\alpha'_R)|}$$

Les Fig. 6 et 7 montrent l'évolution de ces ratios en faisant varier les taux de duplication et transfert et la hauteur de l'arbre. La Fig. 6 montre que $FP_{\mathbb{D}}$ est proche de zéro pour toutes les combinaisons de r_δ, r_τ et de hauteur de l'arbre : quasiment toutes les duplications inférées par notre algorithme sont présentes dans α'_R. Les valeurs très élevées de $FN_{\mathbb{D}}$ peuvent avoir plusieurs causes : α'_R peut contenir des duplications impossible à détecter; $\delta = \tau$ peut amener à inférer un événement de type $\mathbb{T}$ au lieu de $\mathbb{D}$ (ce qui expliquerait aussi le taux très élevé de $FP_{\mathbb{T}}$ de la Fig. 7); la pluralité des réconciliations plus parcimonieuses (cela expliquerait aussi le taux élevé de $FN_{\mathbb{T}}$).

Pour des arbres de 100 espèces et des taux faibles (resp. élevés), l'algorithme résout le MPR avec un temps moyen de 1.09 (resp. 1.38) secondes.

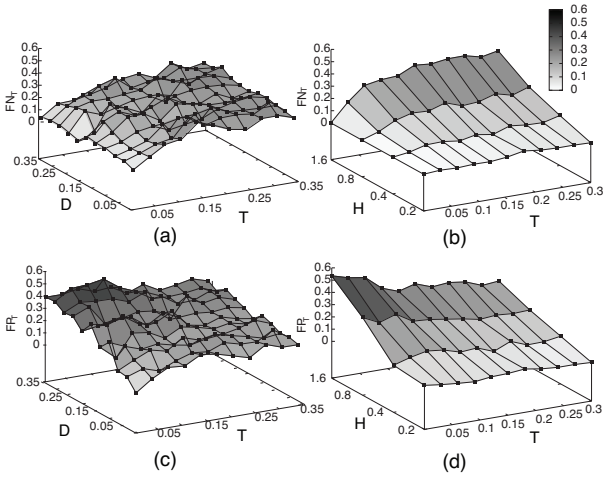


Fig. 7. Étude des conditions dans lesquelles la parcimonie retrouve précisément les événement $\mathbb{DTL}$. Ratios de faux négatifs (a-b) et faux positifs (c-d) pour les événement $\mathbb{T}$ en faisant varier r_δ, r_τ et la hauteur de l'arbre, pour ds_1 (a-c) et ds_2 (b-d).

4 Conclusion

Nos simulations montrent que la parcimonie a de bons résultats sous des conditions réalistes au niveau des phylums. Entre phylums, les transferts sont plus difficiles à retrouver et la pluralité de réconciliations les plus parcimonieuses est probablement un facteur important.

Remerciements

Nous remercions J.-F. Dufayard pour son aide avec les différents programmes de réconciliation, K. Gorbunov et V. Lyubetsky pour les discussions sur [9]. Ce travail est financé par le projet ANR-08-EMER-011 et la Région Languedoc-Roussillon.

Références

[1] L. Arvestad, A. C. Berglund, J. Lagergren, and B. Sennblad. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, 19 Suppl 1 :7–15, 2003.
[2] A. Bernal, U. Ear, and N. Kyrpides. Genomes OnLine Database (GOLD) : a monitor of genome projects world-wide. *Nucleic Acids Res.*, 29 :126–127, 2001.
[3] C. Conow, D. Fielder, Y. Ovadia, and R. Libeskind-Hadas. Jane : a new tool for the cophylogeny reconstruction problem. *Algorithms Mol Biol*, 5 :16, 2010.
[4] M. Csuros and I. Miklos. Streamlining and Large Ancestral Genomes in Archaea Inferred with a Phylogenetic Birth-and-Death Model. *Mol Biol Evol*, 26(9) :2087–2095, 2009.
[5] V. Daubin, N. A. Moran, and H. Ochman. Phylogenetics and the cohesion of bacterial genomes. *Science*, 301 :829–832, 2003.
[6] W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284 :2124–2129, 1999.
[7] N. Goldenfeld and C. Woese. Biology's next revolution. *Nature*, 445 :369, 2007.
[8] M. Goodman, J. Czelusniak, G. W. Moore, Romero A. Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.*, 28 :132–163, 1979.
[9] K. I. Gorbunov and V. A. Lyubetsky. Reconstructing genes evolution along a species tree. *Mol. Biol. (Mosk.)*, 43 :946–958, 2009.
[10] P. Górecki. Reconciliation problems for duplication, loss and horizontal gene transfer. In *RECOMB*, 2004.
[11] R. Guigo, I. Muchnik, and T. F. Smith. Reconstruction of ancient molecular phylogeny. *Mol. Phylogenet. Evol.*, 6 :189–213, 1996.
[12] M. Hallett, J. Lagergren, and A. Tofigh. Simultaneous identification of duplications and lateral transfers. In *RECOMB '04*, pp. 347–356, New York, NY, USA, 2004. San Diego, California, USA, ACM.
[13] C. G. Kurland, B. Canback, and O. G. Berg. Horizontal gene transfer : a critical view. *Proc. Natl. Acad. Sci. U.S.A.*, 100 :9658–9662, 2003.
[14] R. Libeskind-Hadas and M. A. Charleston. On the computational complexity of the reticulate cophylogeny reconstruction problem. *JCB*, 16(1) :105–117, 2009.
[15] S.P. Loader, D. Pisani, J.A. Cotton, D.J. Gower, J.J. Day, and M. Wilkinson. Relative time scales reveal multiple origins of parallel disjunct distributions of african caecilian amphibians. *Biol Lett.*, pp. 505–508, October 2007.
[16] J. O. McInerney, J. A. Cotton, and D. Pisani. The prokaryotic tree of life : past, present... and future ? *Trends Ecol. Evol. (Amst.)*, 23 :276–281, 2008.
[17] D. Merkle and M. Middendorf. Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory Biosci*, 123(4) :277–299, 2005.
[18] D. Merkle, M. Middendorf, and N. Wieseke. A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinformatics*, 11(Suppl 1) :S60, 2010.
[19] R. D. Page. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.*, 43 :58–77, 1994.
[20] S. Penel, A. M. Arigon, J. F. Dufayard, A. S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perriere. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10 Suppl 6 :S3, 2009.
[21] P. Puigbo, Y. Wolf, and E. Koonin. Search for a 'tree of life' in the thicket of the phylogenetic forest. *Journal of Biology*, 8(6) :59, 2009.
[22] A. Rambaut. Phylogen : phylogenetic tree simulator package, 2002.
[23] A. Tofigh. *Using Trees to Capture Reticulate Evolution, Lateral Gene Transfers and Cancer Progression*. PhD thesis, KTH Royal Institute of Technology, Sweden, 2009.
[24] A. Tofigh, M. Hallett, and J. Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM TCBB*, 99, 2010.
[25] A. Tofigh, J. Sjöstrand, B. Sennblad, L. Arvestad, and J. Lagergren. Detecting LGTs using a novel probabilistic model integrating duplications, lgts, losses, rate variation, and sequence evolution, Manuscrit.
[26] B. Vernot, M. Stolzer, A. Goldman, and D. Durand. Reconciliation with non-binary species trees. *J. Comput. Biol.*, 15 :981–1006, 2008.

Exploring the Monochromatic Landscape in Yeast using Genetic Interactions and Known Processes Reveals the Importance of Protein Complexes

Magali MICHAUT[1], Anastasia BARYSHNIKOVA[1], Michael COSTANZO[1], Chad L. MYERS[2], Brenda ANDREWS[1], Charlie BOONE[1] and Gary D. BADER[1]

1 Terrence Donnelly CCBR, University of Toronto, 160 College Street, M5S 3E1, Toronto, Ontario, Canada
{magali.michaut, a.baryshnikova, michael.costanzo, brenda.andrews, charlie.boone, gary.bader}@utoronto.ca

2 Department of Computer Science and Engineering, University of Minnesota, MN 55455, Minneapolis, USA
cmyers@cs.umn.edu

Abstract *If perturbing two genes together has a stronger effect than expected they are said to genetically interact. Recently experimental methods have enabled mapping of positive and negative interactions on a global scale. Negative interactions indicate buffering between genes and positive interactions suggest that the genes are part of the same process. In such networks, some genes interact in a monochromatic manner: interactions between them are mostly positive or negative. It has been proposed that monochromatic gene groups are functionally related and enriched in protein complexes and pathways. Nevertheless system boundaries and relationships are still difficult to define. We evaluate the current model and study the monochromatic nature of biological processes using the most comprehensive quantitative genetic interactions available in the budding yeast Saccharomyces cerevisiae. This new data set includes measurements for 5.4 million pairs of genes and provides quantitative genetic interaction profiles for ~75% of all genes in S. cerevisiae.*

We assess the monochromatic nature of a process as defined in Gene Ontology using a score based on the relative ratio of positive to negative interactions and compute z-scores using a random network model. We also score the monochromatic nature of inter-process connections using a statistical test. We show that 10% of the biological processes are monochromatic and identify 1% of the connections between processes as monochromatic. We then study different features that may explain the monochromaticity and show that protein complexes have a strong contribution. In fact, 63% of the interactions are attributed to complexes whereas we expect only 49%

This work is the first systematic study of the monochromatic nature of biological processes and connections between them. It reveals the importance of protein complexes in the yeast genetic landscape.

Keywords Genetic interaction network, biological process, protein complex, yeast.

1 Introduction

One of the major goals in biology currently is to understand how molecules are organized within the cell and how they interact with each other to perform biological processes. This knowledge can further help to unravel the mechanisms regulating biological processes, why these mechanisms sometimes fail leading to disease and which strategies can help with disease understanding and treatment.

Genetic perturbations are often used in order to better understand the function of a gene and its associated products and to study the relationships between genotype and phenotype [1]. In budding yeast, a commonly used perturbation is gene deletion and a commonly studied phenotype is cell growth. However most yeast gene deletions (~80%) do not affect cell growth in rich medium [1]. To study the function of these genes, two main strategies have been used: i) exploring the phenotypes in different conditions such as the presence of a chemical or an environmental stress [2]; ii) combining mutations in more than one gene to investigate higher-order perturbation effects [3]. This latter approach includes genetic interactions where pairs of mutated genes are tested. Genetic interactions have proven particularly useful to predict gene function [4] and organize biological processes [5] and are complementary to other functional association networks like

protein-protein interactions.

Genetic interactions are observed when the phenotype of a double mutant is unexpected given the phenotypes of both single mutants [6]. Comparing the observed and expected phenotypes, we can classify the genetic interactions into positive and negative. If the phenotype is fitness, a positive (resp. negative) interaction means that the fitness of the double mutant is higher (resp. lower) than expected.

Negative interactions indicate redundancy between two genes, the extreme case occurring when the simultaneous deletion of two non-essential genes is lethal, called 'synthetic lethality'. The biochemical explanation for this is that the two genes participate in complementary or parallel pathways or complexes [7, 8]. As a result, two complementary pathways tend to be connected by many negative interactions. Genetic interactions have thus been used to investigate the organization of the genes into pathways [9].

Positive interactions suggest another type of functional relationships between the genes. In the case of a linear chain of reactions such as a signaling cascade or a biosynthetic pathway, the deletion of one gene or the other would affect the output of the chain. As a result, deleting a second gene is likely not to affect the output more, resulting in a less than expected ('positive') interaction with the first gene. It should be noted that, even if positive interactions indicate that the phenotype of the double mutant is better than expected, it often still results in a decrease in the total phenotypic output. For example, in terms of fitness, a double mutant growing less than the wild type strain, but more than expected based on both single mutant growth rates, would result in a positive interaction. Since the phenotypic output of the double mutant in this case is often between the wild type and the expected output, phenotypic values of each mutant are closer and differences are more difficult to confidently detect. Consequently positive interactions are in general more subtle and difficult to detect than negative interactions [10].

Recently experimental methods have been developed that can be used to measure both positive and negative interactions in a quantitative manner on a genome wide scale - Synthetic Genetic Array (SGA) [5], Epistatic MiniArray (E-MAP) [11] or diploid Synthetic Lethality Analysis by Microarray (dSLAM) [12].

Initial analysis of these experimental data used hierarchical clustering to group functionally related genes [4, 11]. The resulting clusters were manually investigated in order to identify known pathways and complexes and better understand the organization of the early secretory pathway [11], chromatin modifying complexes [13], the homologous recombination pathway [14] and the 26S proteasome [15].

This approach based on hierarchical clustering was extended to consider the types of interactions (positive/negative) between the different clusters. Segre et al. investigated the monochromatic nature of the connections defined as the ratio between positive and negative interactions connecting different clusters [16]. They developed the PRISM algorithm that maximizes the monochromatic nature of the connections (interactions between the clusters are mostly positive or mostly negative) and showed that the results revealed the modular organization of biological systems [16]. This monochromatic pattern also appears within biological processes and in particular within protein complexes [17]. Thus, monochromatic properties of genetic interaction networks can help define modules in the cell and define how they are connected, charting a hierarchical and modular map of the biological systems in the cell.

Multiple methods have been developed to identify pairs of buffering, or complementary, pathways or complexes. These methods are based on the parallel pathway or complex model that involves two groups of genes highly connected to each other by negative interactions. Ma et al. investigated the compensatory properties of biological processes using a graph-based approach on the synthetic lethal network and showed that many cellular functions have genetically compensatory properties by identifying numbers of pairs of buffering pathways [18]. Le Meur et al. found that synthetic lethal interactions can arise from subunits of an essential multi-protein complex or between pairs of multi-protein complexes [19]. Zhang et al. examined the structure of a multi-color network where each color represents a type of interactions (protein-protein interaction, genetic interaction, transcriptional regulation, sequence homology, expression correlation). They found many enriched multi-color network motifs [20].

Another set of methods used protein-protein interaction data to define modules and the connections between them. Modules were defined as clusters of proteins enriched in physical interactions, genetic interactions occurring either within modules (within-pathway model) or between modules (between-pathway model) [17]. This approach was extended by defining modules as a connected graph

in the protein interaction network rather than enriched with protein interactions [21] and considering pairs of complementing modules [22]. Brady et al. introduced stable bipartite subgraphs as a way to identify redundant pathways using synthetic lethal interactions between non-essential genes [23].

On the one hand, negative interactions are thought to occur between buffering pathways. Kelley et al. found that synthetic lethal interactions were better explanations for between-pathways than within-pathways [17]. Nevertheless, negative interactions can also occur within pathways, as is the case for some multi subunit complexes [8]. Moreover complexes enriched in negative interactions tend to contain essential genes [24]. On the other hand, positive interactions were proposed to occur mainly within pathways [13]. The main interpretation of this result is that the deletion of any of the genes in the pathway has an important effect on the pathway activity, which hides the effect of any additional deletion [21], as is the case for a linear cascade.

Most of these methods investigated genetic interactions using only qualitative negative interactions (synthetic lethal interactions) or considering positive and negative interactions as the same type of interactions without any distinction [24]. Some of them are based on unsupervised clustering on the genetic interaction network in order to identify modules [17, 21]. These modules, or systems, are assumed to be complexes or pathways without a clear distinction, but previously identified modules are mainly protein complexes, each defined as a flat list of genes. Some methods are focused on complexes only and trained on reference sets of protein complexes [24]. Thus, the generality of the conclusions about the genetic interactions occurring within and between biological processes is currently not clear and system boundaries are still difficult to define. Even though the monochromatic nature of gene sets has been used to identify biological systems, it is not clear to which extent the different processes are monochromatic or not and to which extent the connections between them are monochromatic.

We propose to use current knowledge about biological processes and to study the monochromatic nature within and between processes using the most comprehensive quantitative genetic interaction data set currently available in the budding yeast that includes measurements for 5.4 million pairs of genes and provides quantitative genetic interaction profiles for ~75% of all genes in *S. cerevisiae* [5].

2 Results and Discussion

2.1 Monochromatic Processes

To study the monochromatic nature of known biological processes, we use the most recent data set of quantitative genetic interactions, obtained using SGA [5]. The known processes are defined by the Gene Ontology (GO) Biological Process (BP) [25] classification system. Each process is defined by a standard name and a set of genes annotated to it. We consider all processes in yeast where these genes are connected by at least one observed SGA interaction (~1000 processes).

We define the monochromatic score as the relative ratio of positive to negative interactions occurring within a given process (set of genes). To assess how likely these scores are to occur by chance, we generate random networks and compute z-scores. We can then identify unexpected patterns by their high z-scores. Highly positive z-scores characterize monochromatic positive processes and highly negative z-scores characterize monochromatic negative processes (see Methods).

Not all genes are tested in the SGA data set, thus processes are variably covered in terms of genetic interactions. We assess the coverage of a process by the number of genes present and connected in the genetic interaction network (see Methods). For a specific level of coverage, we compute the ratio of monochromatic processes among all covered processes. We find that this ratio ranges from 7 to 9% (Tab 1).

Coverage	Covered	Monochromatic	Ratio (%)
0	1031	68	6.6
0.2	1019	68	6.7
0.4	833	66	7.9
0.6	566	50	8.8
0.8	99	9	9.0

Tab. 1. Monochromatic processes covered by SGA.

Choosing a reasonable coverage cut-off of 0.6, we identify 50 monochromatic GO terms, including 5 positive and 45 negative (Tab 2). Positive monochromatic processes are generally much smaller (<= 40 genes) than the negative ones (~100 genes).

Monochromatic processes	Sign
Microautophagy	+
Replication fork processing	+
Histone exchange	+

ER-nuclear signalling pathway	-
Protein transport	-
Cell cycle	-
Reproduction	-
Cell wall organization	-
DNA repair	-

Tab. 2. Examples of monochromatic processes.

Thus, just under 10% of SGA covered biological processes are monochromatic.

2.2 Monochromatic Connections

We then investigated the monochromatic nature of the connections between pairs of biological processes. We consider the set of biological processes previously defined with more than one observed interaction and all possible pairs between these processes. For a given pair of processes, we consider the interactions occurring between two genes from both processes. If some genes are part of both processes, they are not considered. Each connection is thus defined by a set of interactions.

The monochromatic nature of a set of interactions is assessed by a statistical test (Fisher) comparing the number of positive and negative interactions to the background network (see Methods). The p-value is used as a score to select the most monochromatic connections.

The coverage of a connection is assessed by the number of interactions tested among all possible interactions between the two processes. Using different cut-offs on the coverage, we find that ~0.27% of the covered connections are monochromatic (Tab 3). With a cut-off at 0.6 we identify 1386 monochromatic connections, including 613 positive (44%) and 773 negative connections (56%).

Coverage	Covered	Monochromatic	Ratio (%)
0	525727	1394	0.27
0.2	525710	1394	0.27
0.4	525380	1394	0.27
0.6	511671	1386	0.27
0.8	240680	609	0.25

Tab. 3. Monochromatic connections between processes covered by SGA.

Thus, monochromatic connections between pairs of biological processes are rare (~1%).

2.3 Protein Complexes Explain Most Monochromatic Processes

We noticed that the monochromatic processes previously identified often contain protein complexes or part of complexes. All monochromatic processes but six contain at least one gene that is part of a complex. Since complexes tend to be monochromatic (Baryshnikova et al submitted), we evaluated their contribution to the monochromatic patterns we have observed. To address this, we removed the effect of protein complexes and performed the same monochromatic analysis. We removed the effect of complexes in two ways: i) remove all genes that are part of at least one complex; ii) remove the interactions that occur between two genes from the same complex, but leaving the genes in place (in the former case all interactions involving these genes are removed whereas in the latter case only interactions between two genes of the same complex are removed).

When we remove all the genes that are part of a complex, most (82%) of the monochromatic processes identified previously are no longer monochromatic, suggesting that the genes in complexes explain this monochromatic pattern. When we remove the interactions occurring within complexes, only some (28%) of the monochromatic patterns are explained (Tab 3). These results hold for various coverage cut-offs. This indicates that genes which products are part of a protein complex play a key role in the monochromatic patterns identified previously.

We also consider three other features that may contribute to monochromatic patterns: essential genes, duplicate genes and low single mutant fitness genes. Essential genes are known to have many negative interactions [10], duplicate genes often buffer each other and thus are typically connected by a strong negative interaction [26] and genes which have a strong effect on yeast fitness, as measured by growth rate, when deleted (*i.e.* a low single mutant fitness) usually show many negative interactions [10]. Thus, we removed each of these gene sets in turn and evaluate the effect on our observed monochromatic patterns. All these different features partly explain the monochromatic patterns previously identified but not as much as the genes in complex (Tab 4). In addition, these features are highly overlapping with the genes in complex. For example 60% of the essential genes are in a complex. As a result, these features seem to have a minor effect on the monochromatic pattern, which is mainly due to their correlation with the complex feature.

Feature	Genes in complex	Int complex	Essential genes	Duplicate int	Low SMF genes
Processes explained (%)	28	84	54	18	24
Connections explained (%)	67	97	74	56	59

Tab. 4. Monochromatic processes and connections that are explained by removing features.

In summary, we find that proteins in complexes are the most important contributor of monochromatic patterns in a biological process. This suggests that protein complexes play a major role in the monochromatic nature of biological processes.

Protein Complexes Explain Most Monochromatic Connections

In order to assess to which extent the features presented above explain the monochromatic nature of connections, we adopt the same strategy of removing each feature in turn and analyzing the resulting change in monochromatic nature of the connections.

Again, genes encoding proteins which are part of a complex explain most of the monochromatic connections (98%) whereas the other features only partly explain the monochromatic connections (60%). These results hold for various coverage cut-offs. Thus, protein complexes are the most important contributor as was the case for the monochromatic processes. Removing the same number of random genes not in any complex does not have the same effect on the monochromatic pattern. These results suggest that genes in complex play a key role in the monochromatic connections between biological processes.

2.4 30% more of SGA Interactions than Expected are Attributed to Complexes

Following up on the important role of protein complexes in genetically monochromatic processes, we examined the result at the genetic interaction level. If at least one of a pair of genes encodes a protein that is part of a complex, that gene pair is defined as being involved in a protein complex, and otherwise is not. We study both types of gene pairs for the set of all observed interactions in SGA. 49% of all tested pairs of genes (2,801,630 pairs) involve protein complexes and 189,996 interactions were observed. Thus, it is expected that 49% (93,383) interactions should involve a protein complex gene. Surprisingly, we find that 63% (119,871) of the observed SGA genetic interactions involve complexes, or 30% more than expected. This highly significant bias (Fisher $p<10\text{-}5$) is present globally and for both negative and positive interactions considered individually.

To further examine the importance of protein complexes on the topology of the genetic interaction network, we quantified the dependence of the network structure on the following attribute: whether the gene (node) encodes a protein that is part of a complex. We applied a recently published algorithm to assess the importance of node features on network structure [27]. This indicator, based on entropy, has a value higher than 100 for the complex feature in the SGA network. This high number indicates that the 'protein complex' node feature has a strong impact on the topology of the network.

All together these results indicate that genes in complexes are more likely to interact than genes not in complex. This suggests that protein complexes have a disproportionately important role not only in the monochromatic landscape but also more generally in the genetic interaction network in yeast.

3 Conclusions

In this work we study the monochromatic landscape in yeast in a systematic manner using known biological processes as described in GO and a large network of genetic interactions. Approximately 10% of known biological processes, sufficiently covered in terms of interactions, are monochromatic. Only 1% of all pairs of processes interact in a monochromatic manner.

Even though pathways are expected to be monochromatic positive according to the current model, we find many more negative than positive processes. This may be due to the lower experimental detection sensitivity of positive interactions [10] or positive interactions may have a complex mechanistic interpretation [28] requiring the updating of our models.

Considering various features, we showed that protein complexes are extremely important in the monochromatic landscape and the number of interactions is highly biased towards interactions involving complexes. If we describe essential genes as the first level of importance (each gene is individually important), this work suggests that protein complexes can be described as the second

level of importance (pairs of genes are important). We suspect that complexes are more sensitive because they are big machineries and more difficult to buffer, either because it is more difficult to duplicate the functionality of an entire complex or that complexes participate in more processes compared to individual proteins.

We chose GO as the representation of current known biological processes since it is the most comprehensive resource. KEGG and SGD BioCyc also make available non-GO pathway information, but these are limited mostly to metabolic pathways and don't cover as many genes as GO, making a general analysis difficult.

GO organizes processes in a hierarchical structure, which clarifies the relationships between pathways and sub-pathways. However, this makes processes highly overlapping. The number of monochromatic processes depends on this overlap. To assess the effect of overlap, we applied our method on the reduced ontology GO Slim, which contains fewer and less overlapping terms compared to the full GO. We identified 19 monochromatic processes among 36 covered processes. Interestingly, around half of GO Slim processes are monochromatic.

Because of the extensive homology between yeast and human biochemical pathways, such as the cell cycle, it is likely that the yeast cell map will be relevant for improving our understanding of how common Human diseases result from many different possible genotypes composed of many genes. Furthermore the analysis methods we developed is be applicable to other species for which genetic interactions become progressively available such as *C. elegans* [29]. *D. melanogaster* [30] or mammalian cells [31].

4 Material and Methods

4.1 Genetic Interaction Network

The genetic interaction data come from the most recent and comprehensive study in yeast, obtained by the Synthetic Genetic Array technique (SGA) [5]. This data set consists of ~200,000 pair-wise interactions between ~4,400 genes, derived from ~1,700 full genome screens. Each interaction is characterized by the epsilon score, a quantitative genetic interaction measure. This score can be positive or negative, indicating a positive or a negative interaction. When different measurements are available for a single gene (i.e. from several screened alleles of essential genes), we merge all interactions. The resulting network contains 166,401 pair-wise interactions among 4,415 genes.

4.2 Biological Processes

We downloaded the annotation of the yeast genome provided by the Gene Ontology (GO) [25] on September 7th, 2009. For the Biological Process aspect of the ontology, all gene annotations to one specific GO term are up-propagated to all parents of that GO term. We don't consider the annotation coming only from IEA evidence code. We only consider GO terms with more than one observed interaction between its genes and with less than 200 genes in the genetic interaction matrix, otherwise the random networks are not different enough to assess the statistical significance of the monochromatic scores. We thus have a set of 1,031 processes in yeast with genetic interactions in SGA.

4.3 Assessing the Coverage of a GO term

For a given GO term, its genes can be present in the genetic interaction network or not. If present, they contribute to the monochromatic nature only if they are connected within the GO term. We assess the coverage of the GO term by the minimum value of the two following ratios: (i) number of genes in the GO term and in the genetic interaction network over number of genes in the GO term; (ii) number of connected genes in the GO term over number of genes in the GO term and in the genetic interaction network.

4.4 Assessing the Monochromatic Nature of a GO term

We define the monochromatic score of a GO term as the relative ratio of positive to negative interactions observed between the genes in that GO term (see equation 1).

$$(1)\quad S(t) = \frac{\sum_{i \in I} score(i)}{\sum_{i \in I} |score(i)|}$$

where I is the set of interactions occurring between two genes from the GO term t. This score ranges from -1, meaning fully monochromatic negative, to +1, meaning fully monochromatic positive.

We then generate 350 random networks by shuffling the labels of the nodes (the topology is thus conserved). For each GO term, we compute a series

of monochromatic scores obtained with the random genetic interaction networks and use this distribution of scores to compute a z-score (see equation 2).

$$(2)\quad Z = \frac{S - \mu}{\sigma}$$

where S is the monochromatic score to be standardized, μ is the mean of the random scores and σ the standard deviation of the random scores.

4.5 Assessing the Monochromatic Nature of a Connection

A connection between two GO terms is formed by all interactions between one gene belonging to one GO term and another gene belonging to the other GO term. Genes belonging to both GO terms are not considered. We consider the number of positive and negative interactions and test if this ratio follows the background ratio of the whole network (Fisher test). We then select the most significant connections with p-value < 0.01.

4.6 Defining Protein Complexes

We use the cellular component aspect of the Gene Ontology to define protein complexes in yeast. We consider all the children of the GO term macromolecular complex (GO:0032991). Each term defines a protein complex formed by the genes directly annotated to that term (not considering IEA annotations).

4.7 Removing Features

We consider the five following features, removing either genes or interactions: 1) Essential genes: genes described as essential genes in the Saccharomyces Genome Deletion Project [32]; 2) Low SMF genes: genes with low single mutant fitness [5] (10% lowest); 3) Complex genes: genes being part of at least one complex (see the definition of the complexes above); 4) Complex interactions: interactions occurring between two genes being part of at least one complex (see the definition of the complexes above); 5) Intra paralog interactions: interactions occurring between two duplicate genes. The set of duplicate pairs is a combination of the whole genome duplication (WGD) data set from Byrne et al. [33] and smaller-scale duplicates (SSD) [26]. SSD are defined based on sequence similarity with an alignment that covers more than 50% of the length of the longer protein and a BLAST e-value< 10-10.

4.8 Interaction Bias

To examine the role of protein complexes at the interaction level, we study all possible gene pairs. A given pair is involved in a complex if at least one of the genes encodes a protein that is part of a complex. In other words, we partition the genes into two classes: CG, genes that encode a protein that is part of at least one complex; NCG: genes that encode a protein that is not part of any complex. And we partition the interactions into two classes: CI, interactions involving at least one gene of the class CG; NCI, interactions occurring between two genes of the class NCG.

Assuming that the complexes do not have an effect on the structure of the genetic interaction network, we expect the distribution of interaction number among the classes to be the same as the background distribution of all tested pairs. For each interaction class (CI/NCI) we compute the ratio of observed/expected number of interactions.

Acknowledgments

The authors thank Gabriel Musso for his help collecting the paralog dataset.

References

[1] G Giaever, AM Chu, L Ni, C Connelly, L Riles, S Véronneau, S Dow, A Lucau-Danila, K Anderson, B André *et al*, Functional profiling of the Saccharomyces cerevisiae genome. Nature, 418(6896):387-391, 2002

[2] ME Hillenmeyer, E Fung, J Wildenhain, SE Pierce, S Hoon, W Lee, M Proctor, RP St Onge, M Tyers, D Koller *et al*, The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science, 320(5874):362-365, 2008

[3] AH Tong, M Evangelista, AB Parsons, H Xu, GD Bader, N Pagé, M Robinson, S Raghibizadeh, CW Hogue, H Bussey *et al*, Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science, 294(5550):2364-2368, 2001

[4] AHY Tong, G Lesage, GD Bader, H Ding, H Xu, X Xin, J Young, GF Berriz, RL Brost, M Chang *et al*, Global mapping of the yeast genetic interaction network. Science, 303(5659):808-813, 2004

[5] M Costanzo, A Baryshnikova, J Bellay, Y Kim, ED Spear, CS Sevier, H Ding, JLY Koh, K Toufighi, S Mostafavi *et al*, The genetic landscape of a cell. Science, 327(5964):425-431, 2010

[6] R Mani, RP St Onge, JL Hartman, G Giaever, FP Roth, Defining genetic interaction. Proc Natl Acad Sci USA, 105(9):3461-3466, 2008

[7] CL Tucker, S Fields, Lethal combinations. Nat Genet, 35(3):204-205, 2003

[8] C Boone, H Bussey, BJ Andrews, Exploring

genetic interactions and networks with yeast. Nat Rev Genet, 8(6):437-449, 2007
[9] FP Roth, HD Lipshitz, BJ Andrews, Q&A: epistasis. Journal of Biology, 8(4):35, 2009
[10] S Dixon, M Costanzo, A Baryshnikova, B Andrews, C Boone, Systematic Mapping of Genetic Interaction Networks. Annu Rev Genet, 2009
[11] M Schuldiner, SR Collins, NJ Thompson, V Denic, A Bhamidipati, T Punna, J Ihmels, B Andrews, C Boone, JF Greenblatt *et al*, Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. Cell, 123(3):507-519, 2005
[12] X Pan, DS Yuan, S-L Ooi, X Wang, S Sookhai-Mahadeo, P Meluh, JD Boeke, dSLAM analysis of genome-wide genetic interactions in Saccharomyces cerevisiae. Methods, 41(2):206-221, 2007
[13] SR Collins, KM Miller, NL Maas, A Roguev, J Fillingham, CS Chu, M Schuldiner, M Gebbia, J Recht, M Shales *et al*, Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature, 446(7137):806-810, 2007
[14] RP St Onge, R Mani, J Oh, M Proctor, E Fung, RW Davis, C Nislow, FP Roth, G Giaever, Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. Nat Genet, 39(2):199-206, 2007
[15] DK Breslow, DM Cameron, SR Collins, M Schuldiner, J Stewart-Ornstein, HW Newman, S Braun, HD Madhani, NJ Krogan, JS Weissman, A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. Nat Methods, 5(8):711-718, 2008
[16] D Segre, A Deluna, GM Church, R Kishony, Modular epistasis in yeast metabolism. Nat Genet, 37(1):77-83, 2005
[17] R Kelley, T Ideker, Systematic interpretation of genetic interactions using protein networks. Nature Biotechnology, 2005
[18] X Ma, AM Tarone, W Li, Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. PLoS One, 3(4):e1922, 2008
[19] N Le Meur, R Gentleman, Modeling synthetic lethality. Genome Biol, 9(9):R135, 2008
[20] LV Zhang, OD King, SL Wong, DS Goldberg, AHY Tong, G Lesage, B Andrews, H Bussey, C Boone, FP Roth, Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J Biol, 4(2):6, 2005
[21] I Ulitsky, R Shamir, Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks. Mol Syst Biol, 3104, 2007
[22] I Ulitsky, T Shlomi, M Kupiec, R Shamir, From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions. Mol Syst Biol, 4209, 2008
[23] A Brady, K Maxwell, N Daniels, LJ Cowen, Fault tolerance in protein interaction networks: stable bipartite subgraphs and redundant pathways. PLoS One, 4(4):e5364, 2009
[24] S Bandyopadhyay, R Kelley, NJ Krogan, T Ideker, Functional maps of protein complexes from quantitative genetic interaction data. PLoS Comput Biol, 4(4):e1000065, 2008
[25] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig *et al*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25(1):25-29, 2000
[26] G Musso, M Costanzo, M Huangfu, AM Smith, J Paw, BJ San Luis, C Boone, G Giaever, C Nislow, A Emili *et al*, The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. Genome Res, 18(7):1092-1099, 2008
[27] G Bianconi, P Pin, M Marsili, Assessing the relevance of node features for network structure. Proc Natl Acad Sci USA, 2009
[28] X He, W Qian, Z Wang, Y Li, J Zhang, Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. Nat Genet, 42(3):272-276,
[29] RS Kamath, AG Fraser, Y Dong, G Poulin, R Durbin, M Gotta, A Kanapin, N Le Bot, S Moreno, M Sohrmann *et al*, Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature, 421(6920):231-237, 2003
[30] M Boutros, AA Kiger, S Armknecht, K Kerr, M Hild, B Koch, SA Haas, R Paro, N Perrimon, Genome-wide RNAi analysis of growth and viability in Drosophila cells. Science, 303(5659):832-835, 2004
[31] J Moffat, DA Grueneberg, X Yang, SY Kim, AM Kloepfer, G Hinkle, B Piqani, TM Eisenhaure, B Luo, JK Grenier *et al*, A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. Cell, 124(6):1283-1298, 2006
[32] A Baudin, O Ozier-Kalogeropoulos, A Denouel, F Lacroute, C Cullin, A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. Nucleic Acids Res, 21(14):3329-3330, 1993
[33] KP Byrne, KH Wolfe, The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res, 15(10):1456-1461, 2005

Ultra-fast sequence clustering from similarity networks with SiLiX

Vincent MIELE, Simon PENEL, Laurent DURET

Laboratoire Biométrie et Biologie Evolutive, UMR CNRS 5558 - Univ. Lyon 1, F-69622, Villeurbanne, France
{miele,penel,duret}@biomserv.univ-lyon1.fr

Abstract *The number of gene sequences that are available for comparative genomics approaches is increasing extremely quickly. A current challenge is to be able to deal with this huge amount of sequences in order to build families of homologous sequences in a reasonable time. We present a novel method, SiLiX, that reconsiders single linkage clustering with a graph theoretical approach. A parallel version of the algorithms is also presented. As a demonstration of the capability of our method, we clustered more than 3 millions sequences from about 2 billion BLAST hits in 7 minutes. The software package SiLiX is freely available at http://lbbe.univ-lyon1.fr/silix.*

Keywords Homologous sequences, single linkage clustering, graph theory, parallelism.

1 Introduction

Proteins can be naturally classified into families of homologous sequences that derive from a common ancestor. The comparison of homologous sequences and the analysis of their phylogenetic relationships provide very useful information regarding the structure, function and evolution of genes. Thanks to the progress of sequencing projects, this comparative approach can now be applied at the whole genome scale in many different taxa, and several databases have been developed to provide a simple access to collections of multiple sequence alignments and phylogenetic trees [15,22,14]. The building of such phylogenomic databases involves three steps that require important computing resources: 1) compare all proteins between each other to detect sequence similarities, 2) cluster homologous sequences into families (that we will call the *clustering* step) and 3) compute multiple sequence alignments and phylogenetic trees for each family. With the recent progress of sequencing technologies, there is an urgent need to *prepare for the deluge* and hence to develop methods able to deal with a huge quantity of sequences. In this paper, we present a new algorithm for the clustering of homologous sequences, based on single transitive links (*single linkage*). We develop a graph-theoretical model of the dataset which is considered as a *similarity network* where sequences are vertices and similarities are edges [3]. To overcome memory limitations we design an online framework [13] in which we see the edges one at a time to update the families dynamically. This approach enables also an incremental procedure where sequences and similarities are added in the dataset such that it would not be necessary to rebuild the families from scratch. Finally, we adopt a divide-and-conquer strategy [6] to face the quantity of data and design a parallel algorithm of which we analyse the theoretical complexity.

This algorithm presents several advantages over other clustering algorithms: it is extremely fast, it requires only limited memory and it can be run on a parallel architecture - which is essential for ensuring its scalability to large datasets. We implemented this method in a software (called SiLiX for *SIngle LInkage Clustering of Sequences*) and we evaluated its computational performances and scalability on a very large dataset of more than 3 million sequences from the HOGENOM phylogenomic database [14]. SiLiX outperforms other existing software both in terms of speed and of memory requirement. We discuss the interest of SiLiX for the clustering of homologous sequences in huge datasets, possibly in combination with other clustering methods.

2 Methods

2.1 Algorithm framework

Filtering. The principle of the single-linkage clustering is that if sequence A is considered homologous to sequence B, and B homologous to C, then A, B and C are grouped in the same family, whatever the level of similarity between A and C. The choice of the sequence similarity criteria that is used to infer homology is therefore an essential parameter of the single-

linkage clustering approach. Different criteria can be used, separately or in combination (percentage of identity, alignment score or E-value, alignment coverage *i.e.* percentage of the sequence length that is aligned). The choice of these criteria depends on the goal of the clustering (see below for a discussion of the criteria used in the HOGENOM database). The first step of the clustering process therefore consists in analyzing pairwise sequence alignments resulting from the all-against-all comparisons (typically a set of alignments obtained with BLAST [2]) to exclude all pairs that do not meet these sequence similarity criteria. This first step (that we will refer to as the *filtering* step) can be time consuming, but can be easily distributed (see below).

Modelling. Here we consider the following second step: given a list of pairs of similar sequences previously positively filtered, group the sequences into families. We define an undirected graph $G = (C, E_c)$ with the set of vertices C representing sequences and the sets of edges E_c representing similarities between these sequences. We define $n_c = |C|$ and $m_c = |E_c|$. Naturally, finding sequence families consists in computing the connected components of G. In this paper, we want to adress the case of large n_c and m_c and we therefore develop a parsimonious approach in terms of memory use. We want to examine the edges *online* [13] and avoid storing them into a connectivity matrix. Therefore the classical *Depth-first search* algorithm [20] is not adapted.

To record the connected components of G, we only need to store the information of the partition of C into non-overlapping sub-ensembles called *disjoint-sets* and be able to update this information dynamically. When an edge is examined, we need to execute two operations: *find* the name of the set containing each of the two vertices and *union* these sets by merging their vertices. Initialy each vertex is a set by itself. For this purpose we use the *disjoint-sets data structure* [21,1] which is well suited when the graph is discovered edge by edge. This structure allows efficient implementation of the find and union operations by representing each set as a rooted tree. Pratically, the forest composed by all the trees is implemented as an array $parent$ of size n_c. Each element i of a tree has a parent $parent(i)$ such that $parent(r) = r$ if r is the root of the tree. Consequently the underlying problem consists in building and storing only a novel graph $G^* = (C, E_c^*)$, subgraph of G, such that G^* is a spanning star forest: it is actually straightforward and practical to transform each rooted tree into a star tree such that the $parent$ information is a common label for the vertices in a connected component. This will allow to directly retrieve each sequence family by reading the $parent$ information.

Online algorithm for a set of similarities. To build G^* from a set of sequence similarities, we develop a two steps procedure: (1) iteratively build a collection of trees representing the connected components of the graph G and (2) transform each resulting tree into a star tree. For the first step, we adopt the algorithm called *Union-Find by rank with path compression* [21,1]. It consists in updating rooted trees of minimal height while discovering the edges of the graph G online. For this purpose, the $rank$ of a vertex is defined as essentially its height in the tree and each edge (i, j) is processed as explained in Algorithm 1 (see also Figure Fig. 1). It is basically based on the FIND function that associates the root of the tree containing a vertex of interest and the PATHCOMPRESSION function which connects the vertices in a path to the root of a tree. The time complexity was proved to be in our case almost $O(m_c)$ [1]. The second step is straightforward by using PATHCOMPRESSION in $O(n_c)$ time [24]. This procedure requires the storage of n_c $parent$ and n_c $rank$ values such that the memory requirement is $O(n_c)$.

Algorithm 1 ADDEDGE(i, j) by UNION- FIND

Function: FIND(i): returns the root of the tree containing i
Function: PATHCOMPRESSION(i, r): $parent$ of vertices in the path from i to the root of the tree containing i are set to r

1: $r_1 \leftarrow$ FIND(i); $r_2 \leftarrow$ FIND(j)
2: $k \leftarrow \arg\max_{l=1,2} (rank(r_l))$
3: **if** $rank(r_1) == rank(r_2)$ and $r_1 \neq r_2$ **then**
4: $\quad rank(r_k)$++
5: **end if**
6: PATHCOMPRESSION(i, r_k)
7: PATHCOMPRESSION(j, r_k)

Parallelization for multiple sets of similarities. We take advantage of the possibility to explore series of sets of sequence similarities with a client-server parallel architecture. We assume that it is usually affordable to split a large set into q sets. For the sake of clarity, we consider here a group of q processors, which is a reasonable hypothesis in practice. We note that it would also be recommended to have sets of comparable size. We adopt a divide-and-conquer strategy where different processors use the previous sequential algorithm to independently obtain a collection of spanning star forests $G_1^*...G_q^*$ where $G_k^* = (C, E_k^*)$ such that $E_k^* \subset E_c$. These subsolutions are successively merged to obtain the final solution G^* [6]. We first design an algoritm to merge two of these forests in $O(n_c)$ time (see Algorithm 2). It is also based on the disjoint-sets data structure since, for each vertex i, it basically

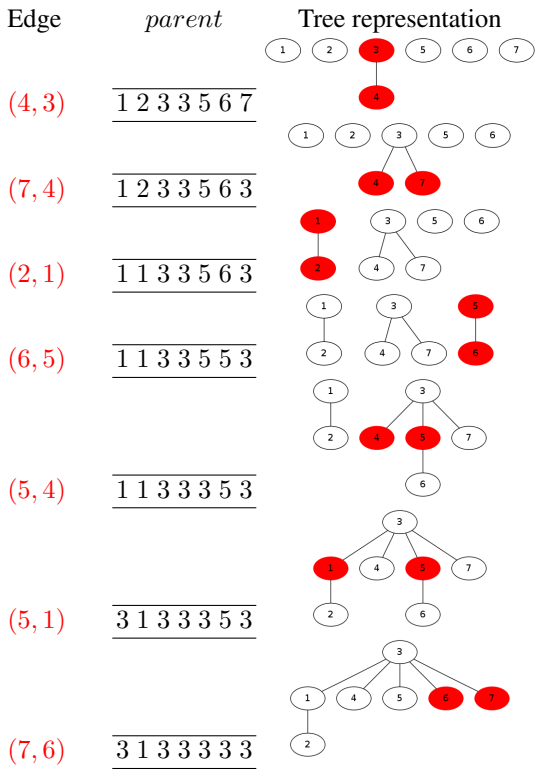


Fig. 1. An example of the steps involved in the algorithm called Union-Find by rank with path compression [21,1]. Edges (first column, in red) are examined online. The disjoint-sets data structure, represented by trees (third column) and implemented using the *parent* array (second column), is consequently modified. The two vertices of the current edge of interest are colored in red.

consists in adding in one forest a formal edge between i and the root of the tree containing i in the other forest. Then we build a parallel formulation of our approach [11,10] where $G_1^*...G_q^*$ are obtained with the step (1) of the sequential algorithm and iteratively merged (see Algorithm 3). The parallel time complexity can be estimated as $O(m_c/q + n_c * q)$. We notice that the merge procedure is many orders of magnitude quicker than the processing of a single set of similarities. For this reason, we decide not to distribute over the processors the merge procedures that will be consequently performed by the server processor in the order of the G_k^* availability.

Algorithm 2 MERGE(G_1^*, G_2^*)

Function: FIND(i): returns the root of the tree containing i

for all i such that FIND(i) $\neq i$ in G_2^* **do**
  r $\leftarrow$ FIND(i) in G_2^*
  ADDEDGE(r, i) in G_1^*
end for

Algorithm 3 Parallel `SiLiX`

each processor r build G_r^* with the sequential algorithm
if $r > 1$ client **then**
  MPI_SEND(G_r^*) to server processor 1
else
  for k in $2, ..., p$ **do**
    MPI_RECEIVE(G_k^*) among $G_2^*, ..., G_q^*$ in their order of availability
    MERGE(G_1^*, G_k^*)
  end for
  for all i **do**
    PATHCOMPRESSION(i, FIND(i))
  end for
end if

2.2 Additional theory for clustering based on alignment coverage with partial sequences

Filtering with partial sequences in the HOGENOM database. HOGENOM is a phylogenomic database of gene families from fully sequenced organisms [14]. The first goal of HOGENOM is to allow the study of the evolution of entire proteins considered as a unit (by contrast with databases such as PFAM [9] or PRODOM [18] that aim at studying the domain architecture of proteins). Hence, in HOGENOM, proteins are classified in the same family only if they are

homologous over their entire length. In practice, protein sequences are compared against each other with BLASTP [2]. For each pairwise alignment, the list of High-scoring Segment Pairs (HSPs) is analyzed to exclude HSPs that are not compatible with a global alignment (for details, see [14]). Then, proteins are classified in the same family if the remaining HSPs cover at least 80% of the longest protein with a percentage of identity greater or equal to 35%. One difficulty is that HOGENOM includes some *partial* protein sequences, because genome sequences are often not 100% complete and hence some genes may overlap with gaps in the genome assembly. These partial sequences cannot be classified using the same criteria as the complete ones and are therefore treated separately. In a first step, gene families are built using only complete protein sequences as explained previously. In a second step, partial sequences are added to this classification, using different alignment length thresholds (for details about parameters, see [14]). It is important to note that, if there are several families that meet these alignment coverage criteria, a partial sequence is included in the one with which it shows the strongest similarity score.

Modelling. To allow the treatment of partial sequences, we propose a modified version of our approach. We define the undirected graph $H = G \cup (P, E_p)$ with two sets of vertices C and P, the complete and partial sequences respectively, and the set of edges E_p between complete and partial sequences, each edge being weighted by the similarity score. We also impose that edges between partial sequences are not allowed. We define $n_p = |P|$, $n = n_c + n_p$, $m_p = |E_p|$ and $m = m_c + m_p$. At this point, we note that sequence families also correspond to connected components but those of a subgraph of H with only the edge of maximum weight conserved for each vertex in P: this will garantee that each partial sequence is connected to only one complete sequence and prevent it to link two connected components. In a similar way than in 2.1, the problem consists in building a novel graph $H^* = G^* \cup (P, E_p^*)$ subgraph of H that has the following properties:

- H^* is a spanning star forest,
- H^* is called a *semi-bipartite graph, i.e.* a graph that can be partitioned into two exclusive and comprehensive parts (C and P) with internal edges (connecting vertices of the same part) only existing within one of the two parts (E_c^*) [4]. The particularity is here that edges between the two parts are weighted,
- $\forall v \in P,\ deg(v) = 1$.

Online algorithm and parallelization. First, it is necessary to insert an additional step between the two steps (1) and (2) of the online algorithm presented in 2.1: build a sub-ensemble of E_p by selecting for each vertex the edge of maximal weight, in $O(m_p)$ time. Then we extend the step (2) to all the vertices in (E, P) for a time complexity in $O(n)$. This procedure runs in $O(n)$ space since it requires the storage of n *parent* values. For the parallelized algorithm, we modify the merging of two forests presented in Algorithm 2 to consider vertices of P and once again select edges of maximal weight, such that the overall parallel complexity can be estimated to be in $O(m/q + n \times q)$.

2.3 Other methods and experimental design

Comparison experiments with other methods were run on a quadri-quadcore Xeon 2.66 GHz with a 12 Gb RAM limit. All programs were run with default parameters. Both programs `SiLiX` and Force [23] take as input all the BLAST hits and perform a first step of filtering then a second step of clustering. MCL [8], MC-UPGMA [12], hcluster_sg [17] and ccomps [7] use a pre-filtered set of couples of sequences IDs with all the partial sequences removed then perform the clustering step. The parallelized version of `SiLiX` was run on a cluster of 2 octo-bicore Opteron 2.8 Ghz and 2 octo-quadcore Opteron 2.3 GHz.

2.4 The `SiLiX` software

All the presented algorithms are implemented into the `SiLiX` software package which is written in ANSI `C++` and uses MPI (Message Passing Interface). `SiLiX` can take two kinds of input. First, the user can provide the result file of an all-against-all BLAST search in tabular format (-m8 option in blastall) in which the diagonal and the upper diagonal hits have been removed (only query-subject hits with query coming after subject in alphabetical order are conserved). In that case, `SiLiX` performs the filtering step by analyzing BLAST hits to search for pairs of sequences that fulfill similarity criteria (alignment coverage, sequence identity) set by the user to build families. In this mode, partial sequences can be treated separately, as described above. Second, if the user prefers to use other types of criteria for the filtering, `SiLiX` can simply take as input a list of pairs of sequences IDs and perform the clustering step. Compilation and installation are compliant with the GNU standard procedure. The library is freely available on the `SiLiX` webpage `http://lbbe.univ-lyon1.fr/silix`. Online documentation and man pages are also available. `SiLiX` is licensed

under the General Public License (http://www.gnu.org/licenses/licenses.html).

3 Results

3.1 SiLiX faster and more memory efficient than other methods

To test SiLiX and compare it to state-of-the-art programs, we extracted protein sequences from the HOGENOM database (Release 5, December 2009, [14]). The current release of HOGENOM contains 820 bacteria, 62 archaea and 51 eukaryotes for 3,666,568 protein sequences (76% bacteria, 3% archae and 20% eukarya). We selected 3,159,593 non-redundant sequences including about 1% partial sequences. Sequences were compared between each other with BLASTP [2] with an E-value threshold set to 10^{-4}. The BLAST output file contains 1,905,335,339 pairwise alignments. We tested five previously published programs, for which the source code is publicly available. These programs can be divided in two categories: those performing the filtering of BLAST hits and the clustering and those performing only the clustering. For a fair comparison, we ran SiLiX in the two modes (filtering+clustering or clustering only). The clustering of the protein dataset with SiLiX is extremely fast (about 3 min) and requires only limited memory capacity (see Table Tab. 1). Interestingly, the filtering of the BLAST result file takes much more time than the clustering itself (see the running times of SiLiX with or without the filtering step). The run time is indeed penalized by the necessity to retrieve the sequence lengths in a yet efficient hash map structure. Meanwhile it is necessary to note that, after the filtering step, the number of similarity pairs given as input to the clustering step represents less than 10% of the number of pairwise alignments. Two of the five other programs (Force [23] and ccomps [7]) turned out to be limited by the available RAM memory (12 Gb) and failed to cluster the protein dataset. MC-UPGMA [12] which is very efficient in terms of memory usage takes an order of magnitude more time than SiLiX. Lastly, hcluster_sg [17] and MCL [8] deal with the dataset in respectively 15x and 60x more time than SiLiX and with high memory requirement. Consequently, SiLiX presents the best abilities to tackle the challenge of huge dataset analysis with CPU and memory requirements equivalent to those of a laptop computer.

3.2 SiLiX scalable in practice

One could be satisfied to be able to deal with a huge dataset in a couple of hours. Meanwhile, the number of available sequences increases dramatically and the number of similarities is quadratic with this number of sequences. Moreover it could be valuable to offer the possibility to run SiLiX with different values of the filtering parameters, to perform a sensitivity analysis for example. For this first reason, we propose to face the need for scalability into a parallel framework. We designed a parallel implementation of SiLiX with a low number of inter-processors communications to take advantage of multiple kinds of parallel hardware architectures. This algorithm delocalizes the processing of the sequence similarity dataset, including the filtering step, and merges the results in a last step (see Methods). We designed a divide and conquer approach that requires only $q-1$ communications where q is the number of processors, followed by merge procedures between partial results from two processors that are considerably faster than the independant computations on each processor. For these reasons, we observe pratical performances consistent with the theoretical complexity such that the run time decrease is inversely-proportional to the number of processors (see Figure Fig. 2).

4 Discussion

Different methods have been proposed for the clustering of proteins into families of homologous sequences [8,12,23,22,15,14,17]. These methods differ both in terms of the quality of the clustering, and in terms of the computing resources necessary to perform the clustering. The single-linkage clustering approach is used in different phylogenomic databases such as EnsemblCompara [22] or HOGENOM [14]. Here we propose a new implementation of the single linkage clustering method, SiLiX, which is extremely efficient both in terms of the computing time and memory requirement. Moreover, this method can be cost-effectively run on parallel architectures, and hence is easily scalable. Thus, in terms of the computing resource requirements, this method is much more efficient than other available methods for the treatment of huge sequence datasets. We do not claim however that SiLiX outperforms other methods in terms of the quality of the clustering. In fact it is known that the single linkage clustering approach can be problematic, because one single false positive link can lead to the clustering of non-homologous sequences in a same family. This risk of false positive links increases with the size of the dataset, and hence the quality of the clustering is expected to decrease as the amount of sequences increases. Thus, the use of SiLiX alone with a very large sequence dataset is likely to give some heterogenous families. However, given its speed,

method	filtering ($> 1.9 \times 10^9$ pairs)	clustering ($> 1.38 \times 10^8$ pairs)	CPU (min)	MEM (Gb)
SiLiX	x	x	138	0.36
Force [23]	x	x	-	Out of Memory
SiLiX		x	3.2	0.24
hcluster_sg [17]		x	51	4.5
MCL [8]		x	194	6
MC-UPGMA [12]		x	617	1.7
ccomps [7]		x	-	Out of Memory

Tab. 1. CPU time and memory requirements for SiLiX and five state-of-the-art programs divided in two categories, those performing the filtering of BLAST hits and the clustering and those performing only the clustering, on the dataset of similarity pairs extracted from the HOGENOM database [14]. After the filtering step, the number of similarity pairs is less than 10% of the original number of pairwise alignments.

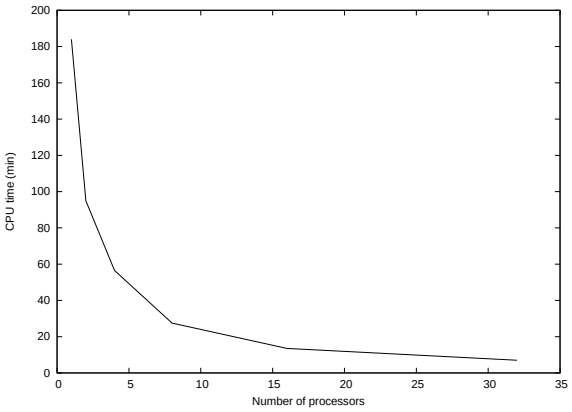


Fig. 2. CPU time of the parallelized version of SiLiX for varying number of processors on the dataset of similarity pairs extracted from the HOGENOM database [14].

SiLiX can efficiently be used as a first clustering step, before running other algorithms. For instance, studying the similarity network of each family from a topological point of view [3] is already affordable with state of the art methods for connectivity or community structure detection ([5,16], see Figure Fig. 3). This would allow to post-treat and curate the families obtained with SiLiX and automatically remove similarities that must be artifactual to consider subclusters inside families. Interesting perspectives could also consist in interpreting topology from an evolutionary perspective.

Acknowledgments

The authors would like to thank D.Kahn, V.Lacroix, M.F. Sagot and E.Tannier for helpful discussions and comments, B.Spataro for the computing facilities and Y.Loewenstein, J.Baumbach and A.Krause for their answers about questions on the availability and use of their programs.
This work has been supported by the French Agence Nationale de la Recherche under grant NeMo ANR-08-BLAN-0304-01.

References

[1] M. H. Alsuwaiyel. *Algorithms: Design Techniques and Analysis*. World Scientific Publishing Company, 1998.

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, Sep 1997.

[3] H. J. Atkinson, J. H. Morris, T. E. Ferrin, and P. C. Babbitt. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE*, 4:e4345, 2009.

[4] Y. Bramoulle, D. Lopez-Pintado, S. Goyal, and F. Vega-Redondo. Network formation and anti-coordination games. *International Journal of Game Theory*, 33(1):1–19, 2004.

[5] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70:066111, Dec 2004.

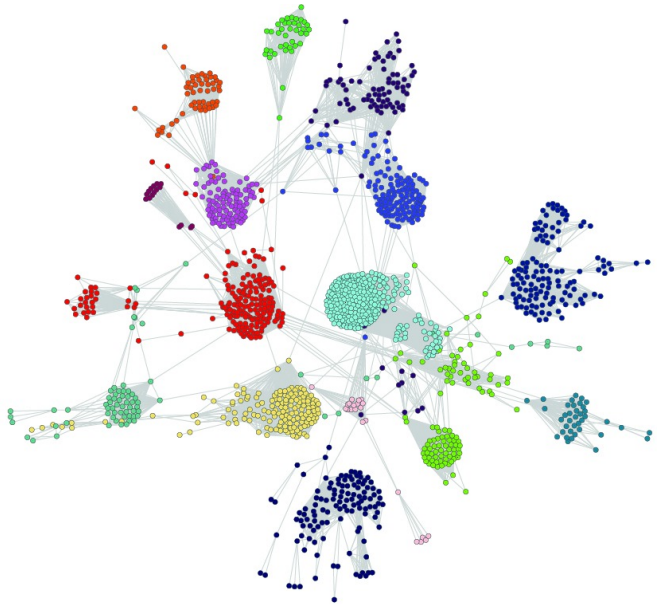


Fig. 3. Similarity network of 1521 complete sequences (vertices) forming a family retrieved by `SiLiX`. Each color corresponds to a community of sequences found by the algorithm of [5]. Representation with Cytoscape [19] with organic layout.

[6] S. K. Das and D. Narsingh. Divide- and- conquer-based optimal parallel algorithms for some graph problems on EREW PRAM model. *IEEE transactions on circuits and systems*, 35(3):312–322, 1988.

[7] J. Ellson, E. Gansner, L. Koutsofios, S. North, and G. Woodhull. Graphviz - open source graph drawing tools. *Lecture notes in computer science*, 2001.

[8] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30:1575–1584, Apr 2002.

[9] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res.*, 38:D211–222, Jan 2010.

[10] Y. Han and R. A. Wagner. An efficient and fast parallel-connected component algorithm. *Journal of the ACM*, 37(3):626–642, 1990.

[11] A. Krishnamurthy, S. S. Lumetta, D. E. Culler, and K. Yelick. Connected Components on Distributed Memory Machines. *Parallel Algorithms, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 1997.

[12] Y. Loewenstein, E. Portugaly, M. Fromer, and M. Linial. Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics*, 24:i41–49, Jul 2008.

[13] D. J. Pearce and P. H. J. Kelly. Online algorithms for topological order and strongly connected components. Technical report, Imperial College, London, 2003.

[14] S. Penel, A. M. Arigon, J. F. Dufayard, A. S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perriere. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10 Suppl 6:S3, 2009.

[15] R. Petryszak, E. Kretschmann, D. Wieser, and R. Apweiler. The predictive power of the CluSTr database. *Bioinformatics*, 21:3604–3609, Sep 2005.

[16] F. Picard, V. Miele, J. J. Daudin, L. Cottret, and S. Robin. Deciphering the connectivity structure of biological networks using MixNet. *BMC Bioinformatics*, 10 Suppl 6:S17, 2009.

[17] J. Ruan, H. Li, Z. Chen, A. Coghlan, L. J. Coin, Y. Guo, J. K. Heriche, Y. Hu, K. Kristiansen, R. Li, T. Liu, A. Moses, J. Qin, S. Vang, A. J. Vilella, A. Ureta-Vidal, L. Bolund, J. Wang, and R. Durbin. TreeFam: 2008 Update. *Nucleic Acids Res.*, 36:D735–740, Jan 2008.

[18] F. Servant, C. Bru, S. Carrere, E. Courcelle, J. Gouzy, D. Peyruc, and D. Kahn. ProDom: automated clustering of homologous domains. *Brief. Bioinformatics*, 3:246–251, Sep 2002.

[19] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13:2498–2504, Nov 2003.

[20] R. Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972.

[21] R. E. Tarjan. Efficiency of a Good But Not Linear Set Union Algorithm. *Journal of the ACM*, 22(2):215–225, 1975.

[22] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic

trees in vertebrates. *Genome Res.*, 19:327–335, Feb 2009.

[23] T. Wittkop, J. Baumbach, F. P. Lobo, and S. Rahmann. Large scale clustering of protein sequences with FORCE -A layout based heuristic for weighted cluster editing. *BMC Bioinformatics*, 8:396, 2007.

[24] K. Wu and E. Otoo. A simpler proof of the average case complexity of union-find with path compression. Technical report, Lawrence Berkeley National Laboratory, Berkeley, 2005.

Piecewise smooth hybrid systems as models for networks in molecular biology

Vincent NOEL[1], Sergei VAKULENKO[4] and Ovidiu RADULESCU[2,3]

[1] Université de Rennes 1 - CNRS 6025 (IRMAR), Campus de Beaulieu, 35042 Rennes, France
vincent.noel@univ-rennes1.fr
[2] Université de Montpellier 2, DIMNP - UMR 5235 CNRS/UM1/UM2, Pl. E. Bataillon, Bat 24, CP 107, 34095 Montpellier Cedex 5, France
[3] INRIA Rennes Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes, France
ovidiu.radulescu@univ-montp2.fr
[4] Saint Petersburg State University of Technology and Design, Bolshoya Morskaya 18, Saint Petersburg, Russia

Abstract *We discuss piecewise smooth hybrid systems as models for regulatory networks in molecular biology. These systems involve both continuous and discrete variables. In the context of gene networks, the discrete variables allow to switch on and off some of the molecular interactions in the model of the biological system. Piecewise smooth hybrid models are well adapted to approximate the dynamics of multiscale dissipative systems that occur in molecular biology. We show how to produce such models by a top down approach that use biological knowledge for a guided choice of important variables and interactions. Then we propose an algorithm for fitting parameters of the piecewise smooth models from data. We illustrate some of the possibilities of this approach by proposing a minimal piecewise smooth model for the cell cycle.*

Keywords systems biology, hybrid models, cell cycle

1 Introduction

Hybrid systems are widely used in automatic control theory to cope with situations arising when a finite-state machine is coupled to mechanisms that can be modeled by differential equations [11]. It is the case of robots, plant controllers, computer disk drives, automated highway systems, flight control, etc. The general behavior of such systems is to pass from one type of smooth dynamics (mode) described by one set of differential equations to another smooth dynamics (mode) described by another set of differential equations. The command of the modes can be performed by changing one or several discrete variables. The mode change can be accompanied or not by jumps (discontinuities) of the trajectories.

Depending on how the discrete variables are changed there may be several types of hybrid systems: switched systems [14], multivalued differential automata [15], piecewise smooth systems [2]. Notice that in the last case, the mode changes when the trajectory attains some smooth manifolds.

Piecewise affine hybrid systems have been used to model dynamics of gene networks [1,3]. In these networks, most of the time, the gene variables are close to discrete values (attractors) and the transitions between discrete attractors are dictated by the relative position of the transient values of these variables with respect to some thresholds. The transient dynamics leading to attractors is considered to be piecewise affine where the linear part of the dynamical equations is defined by a diagonal matrix with negative entries. This approximation allows to reduce the dynamics of simple genetic circuits to a discrete automaton, and can be used for various application such as model checking. However, the study of large networks with this approach suffers from combinatorial explosion.

We must emphasize that piecewise affine models are not always good approximations for the dynamics of the modes. The machinery of the cell cycle is an example. Proteolytic degradation of the cyclins is switched on rapidly by the cyclin dependent kinase complexes but between two successive switchings the complexes have non-linear dynamics implying several positive (autocatalytic processes) and negative feedback loops. These non-linear processes contribute to the robustness of the mechanism. Another example is the dynamics of the genetically regulated metabolism. Genetic changes could be considered as boolean variables that are turned on and off by their mutual interaction and by the interaction with the metabolites, but between two successive switchings of the gene expression the dynamics of metabolism is not linear. More generally, the dynamics of multi-scale network belongs to a patchy landscape formed by smooth, low

dimensional, but curved manifolds, connected by discontinuous transitions. The patches represent low dimensional local invariant manifolds, typical for multiscale dissipative systems, and the transitions correspond to bifurcations of these manifolds [7,6]. Piecewise smooth systems can provide more realistic and more robust models describing these situations.

The idea of piecewise smooth patchy landscape arises naturally from the model reduction theory. The dynamics of a multiscale, but nonlinear large model, can be reduced to the one of a dominant subsystem [12,9,8]. In dynamical systems with separation of timescales the dominant subsystem depends on the relative contributions of different variables to the timescales and on the comparison between timescales. Both the contributions of different variables to the timescales of the dynamics and the comparison among timescales (which timescale is slower which one is quicker) can change along a trajectory of the system. Considering that the set of dominant subsystems is finite, the changes are necessarily discrete. Thus, although one may try and sometimes succeed to find a global reduced model, the general picture in the case of multiscale non-linear dissipative systems is a sequence of several approximations (modes) valid locally. The modes integrate the degrees of freedom of the system that are active for a certain time interval [12,9,8].

The problem of how the modes can be rigorously approximated for a given multiscale nonlinear model will be approached elsewhere. In this paper we propose a heuristic to construct appropriate modes and adequate piecewise smooth models by using a top-down approach. Then, we show how the parameters of the hybrid model can be fitted from data or from trajectories produced by existing smooth, but more complex models.

2 Hybrid models

We consider the so-called hybrid dynamical systems (HDS) consisting of two components: a continuous part, u, defined by

$$\frac{du_i}{dt} = f_i(u(t), s(t)), \quad t > 0, \tag{2.1}$$

where $u = (u_1, u_2, ..., u_n) \in \mathbf{R}^n$, and a discrete part $s(t) \in S$, where S is a finite set of states. For molecular networks, the continuous variables are protein concentrations and the discrete states may be gene activities described by boolean variables $s = (s_1(t), s_2(t), ..., s_m(t))$, where $s_j \in \{0, 1\}$ (such boolean gene models are popular, see [4,10] among many others).

There are several possible ways to define the evolution of the s variables. Rather generally, this can be done by a time continuous Markov chain with transition probabilities $p(s, s', u)$ from the state s to the state s' (per unit time) depending on current state $u(t)$. However, in gene networks, transition probabilities dependence on u is not smooth. For instance, the probability for s to jump is close to one if u goes above some threshold value, and close to zero if u is smaller than the threshold. We can, in certain cases, neglect the transition time with respect to the time needed for u variables to change. Assuming that some of the discrete variables contribute to production of u and that other contribute to the degradation of u we obtain a general model of hybrid piece-wise smooth dynamical system :

$$\begin{aligned}
\frac{du_i}{dt} &= \sum_{k=1}^{N} s_k P_{ik}(u) + P_i^0(u) \\
&\quad - \sum_{l=1}^{M} \tilde{s}_l Q_{il}(u) - Q_i^0(u), \\
s_j &= H(\sum_{k=1}^{n} w_{jk} u_k - h_j), \\
\tilde{s}_l &= H(\sum_{k=1}^{n} \tilde{w}_{lk} u_k - \tilde{h}_l)
\end{aligned} \tag{2.2}$$

where H is the unit step function $H(y) = 1$, $y \geq 0$, and $H(y) = 0$, $y < 0$, $P_{ik}, P_i^0, Q_{il}, Q_i^0$ are positive, smooth functions of u_i representing production, basal production, consumption, and basal consumption, respectively. Here $w, \tilde{w}$ are matrices describing the interactions between the u variables, $i = 1, 2, ..., n$, $j = 1, 2, ..., N$, $l = 1, ..., M$ and $h, \tilde{h}$ are thresholds.

The class of models (2.2) is still too general. We shall restrict ourselves to a subclass of piecewise smooth systems where smooth production and degradation terms will be assumed multivariate monomials in u, plus some basal terms:

$$\begin{aligned}
P_{ik}(\mathbf{u}) &= a_{ik} u_1^{\alpha_1^{ik}} \dots u_n^{\alpha_n^{ik}}, \\
P_i^0(\mathbf{u}) &= a_i^0 \\
Q_{il}(\mathbf{u}) &= \tilde{a}_{il} u_1^{\tilde{\alpha}_1^{il}} \dots u_n^{\tilde{\alpha}_n^{il}} \\
Q_i^0(\mathbf{u}) &= \tilde{a}_i^0 u_i
\end{aligned} \tag{2.3}$$

which will be chosen according to a heuristic presented in the next section.

These models have several advantages with respect to standard models in molecular biology and neuroscience based on differential equations. They allow

us to simulate, in a fairly simple manner, discontinuous transitions occurring in such systems (see a typical graph describing time evolution of protein concentration within cellular cell cycle, Fig. 4.1). The discontinuous transitions result either from fast processes or from strongly non-linear (thresholding) phenomena. This class of models is also scalable in the sense that more and more details can be introduced at relatively low cost, by increasing the number of discrete variables and the size of the interaction matrices.

The definition of the modes slightly extends the one of S-systems, introduced by Savageau [13]. Our choice was motivated by the fact that S-systems proved their utility as models for metabolic networks whose dynamics we want to encompass by considering the modes. The introduction of basal terms avoids spurious long living states when some products have zero concentrations.

The monomial rates can be fully justified for linear networks of biochemical reactions with totally separated constants. The same is true for nonlinear mechanisms resulting from mass action law for instance. In general simplified rates of complex mechanisms can be rational functions of the concentrations. However, when concentrations are very large or very small the monomial power laws are recovered. For a multiscale system changing regime (for instance a Michaelis Menten reaction switching from a saturated enzyme regime to a small concentration substrate regime) one can use the discrete variables to illustrate the change.

In the next section we illustrate the possibilities of this model and show that (2.2) can simulate the mitotic oscillations of the cell cycle.

3 Heuristic for choosing the discrete variables and the multivariate monomial terms

The interactions between the molecular variables of the model can occur at several levels:

i) The discrete interactions.

Discrete interactions manifest themselves punctually as a consequence of thresholding of rapid phenomena. They contribute to changing the discrete variables s_j, $\tilde{s}_j$.

One protein can contribute to switching on or off the discrete variables commanding the production or the degradation of another protein. The action of u_i on u_j is positive (an activation) if $w_{ji} > 0$ (contribute to turn on production) or if $\tilde{w}_{ji} < 0$ (contribute to turn off degradation). Conversely the action of u_i on u_j is negative if $w_{ji} < 0$ or if $\tilde{w}_{ji} > 0$.

ii The continuous interactions.

The continuous interactions guide the dynamics of the modes. During the mode dynamics the variables s_j, $\tilde{s}_j$ are fixed. The continuous variable u_i activates u_j if either $\alpha_j^{ik} > 0$ or $\tilde{\alpha}_j^{il} < 0$, for some k, l. Conversely, u_i inactivates u_j if either $\alpha_j^{ik} < 0$ or $\tilde{\alpha}_j^{il} > 0$, for some k, l.

In the following we provide a heuristic allowing to produce hybrid models.

In order to define a hybrid model we need a hybrid interaction scheme. This consists in saying, for each given species, whether its production/degradation can be switched on and off and by which species, also which species modulate the production/degradation of a given species in a smooth way. The representation of the hybrid interaction scheme can be given as a regulated reaction graph.

A regulated reaction graph is a quadruple (V, R, E, E_r). The triplet (V, R, E), where $E \subset V \times R \cup R \times V$, defines a reaction bipartite graph, ie $(x, y) \in E$ iff $x \in V, y \in R$ and x is a substrate of R, or $x \in R, y \in V$ and y is a product of x. $E_r \subset V \times R$ is the set that defines regulations, for instance $(x, z) \in E_r$ if $x \in V$ regulates $z \in R$.

Consistently with the choice (2.2),(2.3) for piecewise-smooth systems the stoichiometry of the reaction graph (V, R, E) is mono-molecular, any reaction has at most one substrate and at most one product (generalizations are possible, but will not be discussed here).

Some of the regulations in E_r are discrete and some are continuous and we can define the partition $E_r = E_r^d \cup E_r^c$. Similarly, there is a partition of the reactions $R = R^c \cup R^s$. A reaction y belongs to the switched reactions $y \in R^s$ if $(x, y) \in E_r^d$, for some $x \in V$.

The role of the regulators (continuous if they modulate the reaction rate, discrete if they contribute to switching it on and off) should be indicated on the graph together with the signs of the regulations.

Given a reaction, we identify its substrate and the regulators. The non-basal term in the reaction rate is a product of the concentrations of the substrates, concentrations of activators, divided by the concentrations of inhibitors. The basal term is constant if there is no substrate, or proportional to the concentration of the substrate (for instance in consumption reactions).

Assuming that there are n species $u \in \mathbb{R}^n$ and that the reactions have stoichiometric vectors ν_j, $1 \leq j \leq m$, one obtains the following piecewise-smooth model:

$$\frac{d\mathbf{u}}{dt} = \sum_{j\in R^c} \nu_j R_j(\mathbf{u}) + \sum_{k\in R^d} \nu_k (R_k(\mathbf{u})\sigma_k(\mathbf{u}) + R_k^0(\mathbf{u})) \tag{3.1}$$

where $\sigma_k(\mathbf{u}) = H(\sum_{(i,j)\in E^r} w_{kj}u_j - h_k)$. The relation between σ_k and $s_i, \tilde{s}_j$ from Eq.2.2 is straightforward.

The reaction rates have the forms given by (2.3). The monomial exponents α_{ij}, $\tilde{\alpha}_j^i$ and the final rates can be obtain from the following heuristic rules:

i) If a reaction j is activated then $\alpha_j^i = 1$ for all activators and $\alpha_j^i = -1$ for all inhibitors i in the absence of cooperativity. Cooperativity may be taken into account by considering $|\alpha_j^i| > 1$.

ii) Basal rates are constant for reactions without substrates and proportional to the concentration of the substrate otherwise.

iii) If activated reactions are present with intermittence, their non-basal rates are multiplied by discrete variables s_i.

As an example let us consider the minimal model proposed by Goldbeter for mitotic oscillations of the cell cycle [5]. Basically, this consists of three variables C (cyclin), M (cyclin dependent kinase complex) and X (proteolytic enzyme, most probably a polo-like kinase). The production of M is activated by C (also by M which is auto-catalytic), the production of X is activated by M and the degradation of C is activated by X. The hybrid interaction scheme contains six reactions. We decided that the degradation of the cyclin C acts discretely (on/off mechanism) and that all the other reactions are always present in the model (their rates are smoothly regulated). Then the hybrid model is the following:

$$\begin{aligned} \frac{dC}{dt} &= k_1 - \tilde{k}_1 C H(X - \tilde{h}_1) - \tilde{k}_1^0 C \\ \frac{dM}{dt} &= (k_2 MC + k_1^0) - \tilde{k}_2^0 M \\ \frac{dX}{dt} &= (k_3 M + k_2^0) - \tilde{k}_3^0 X \end{aligned} \tag{3.2}$$

where H is the Heaviside unit step function.

4 Reverse engineering of hybrid models

We would like to develop a method allowing to find the parameters of a model from the class introduced above that best describes the observed trajectories of a biological system. These trajectories can come from experiments or can be produced by non-hybrid models. In both situations we obtain a model whose parameters can be easily interpreted in biological terms. The hybrid model can be further analyzed or used to model more complex situations.

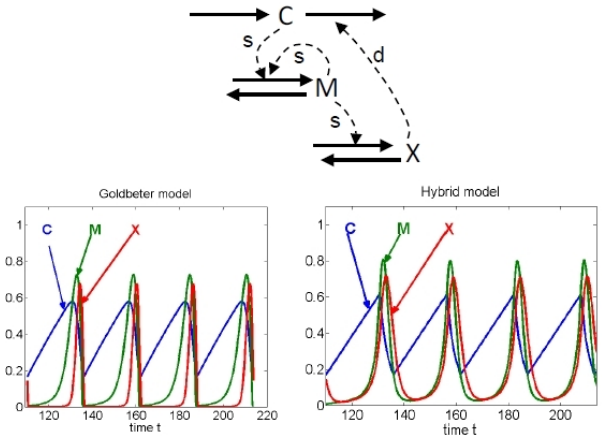


Fig. 4.1. (Middle) Regulated reaction graph for the minimal cell cycle model. Continuous arrows represent reactions, dotted arrows represent regulations. (s) regulations smoothly modulate the rates. (d) regulations discretely turn on and off the reaction rates. (Left) Trajectories of the non-hybrid model by Golbeter [5]. (Right) Trajectories of the hybrid model.

In the following we present a reverse engineering algorithm that works well for systems with sharp transitions.

Data. n trajectories (time series) $u_1(t), ..., u_n(t)$ given at time moments $t_0, t_1, ..., t_N$. A regulated reaction graph (the smooth/discrete partition of the regulations can be unspecified).

Output. A model of the type (2.2),(2.3) with values of the parameters that fit well the data.

The algorithm has several steps.

I. Splitting of the trajectory into smooth parts.

We look for K time intervals $I_1, I_2, ..., I_K$. The dynamics on each of the intervals is smooth, it is given by (2.2) with the s variables fixed. Mode transitions (change of the variables) occur at the borders of these intervals. We denote the switching times as $\tau_1, ...\tau_K$.

Finding τ_k is a problem of singularity detection. This could be done by various methods, for example by wavelet analysis. We have chosen as criterium the value of the second derivative of u_i. For piecewise smooth systems, the derivatives of the trajectories are discontinuous at the switching times τ_k. The second derivative has delta-Dirac components located at τ_k, which will show up as peaks in the numerically estimated second derivatives.

II. Identification of the mode transitions.

Given a switching time τ_k the mode transition is defined by the set of values values σ_j indicating reactions to be turned on or off at τ_k. The presence of a discontinuity is indicated by a peak in the second derivative of one or several species u_i. Without knowing which reaction in the regulated reaction graph has discrete behavior, there are several possible choices for such reactions. Each one of this choices could lead to a different hybrid model corresponding to a different characterization of the interactions as discrete and continuous. This step is supervised and could take into account biologist's intuition.

The discontinuities of the trajectories give the transitions but not the first mode. This choice is also supervised and takes into account periodicity constraints. From the first mode and from the transitions, all the modes (values of σ_j on the intervals I_k) are straightforwardly obtained.

III. Determining the mode internal parameters.

The mode internal parameters are obtained by simulating annealing. Let $u_i^{modes}(t)$ be the continuous hybrid trajectories obtained by integrating the modes between the calculated transition times. The simulated annealing algorithm minimizes the following objective function:

$$F = \sum_{i,k} C_k (u_i^{modes}(t_k) - u_i(t_k))^2$$

C_k are positive weights that increase with time. We thus penalize large time deviations that can arise from period misfit.

IV. Determining the mode control parameters.

Let $\sigma_m = H(\sum_{(m,j)\in E^r} w_{mj} u_j - h_j)$ be the discrete variables determined above. Let σ_k^m be the constant values of σ_m on T_k. Consider now the optimal trajectories $u_i^{modes*}(t_l)$ obtained before.

Then, one should have

$$(\sum_{(m,j)\in E^r} w_{mj} u_j^{modes*}(t_l) - h_j)\sigma_k^m > 0, \text{for all } t_l \in T_k \tag{4.1}$$

which is a linear programming problem for w_{mj} that can be resolved (if it has a solution) in polynomial time.

The algorithm has been applied to the minimal cell cycle model by Golbeter and the result is shown in Fig. 4.1. Of course the fit is not perfect and one should by no means expect a perfect fit. One of the reason of the differences is that the model by Goldbeter uses degradation terms that saturate and are practically constant on the descending slope of the variables M, X, while our linear degradation terms lead to exponential decrease.

5 Conclusion

The results that we present are a proof of principle that piecewise smooth hybrid models can be constructed with a simple heuristic from basic information about biochemical interactions. Using this class of hybrid models instead of piecewise-linear approximations provides, in many situations, a better balance between discrete and smooth interactions. For instance, the hybrid cell cycle model presented here has only two discrete transitions per period and it is very robust. A piecewise-linear version of the same model, would need a lot more discrete transitions per period which will reduce robustness and increase the difficulty of the inversion procedure. In the future we will apply the heuristic and the fitting algorithm to obtain a realistic model for the eucaryotic cell cycle.

References

[1] H. De Jong, J.L. Gouzé, C. Hernandez, M. Page, T. Sari, and J. Geiselmann. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bulletin of Mathematical Biology*, 66(2):301–340, 2004.

[2] A.F. Filippov and FM Arscott. *Differential equations with discontinuous righthand sides*. Springer, 1988.

[3] J. Gebert, N. Radde, and G.W. Weber. Modeling gene regulatory networks with piecewise linear differential equations. *European Journal of Operational Research*, 181(3):1148–1165, 2007.

[4] L. Glass and S.A. Kauffman. The logical analysis of continuous, non-linear biochemical control networks. *Journal of Theoretical Biology*, 39(1):103–129, 1973.

[5] A. Goldbeter. A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *Proceedings of the National Academy of Sciences of the United States of America*, 88(20):9107, 1991.

[6] A.N. Gorban and I.V. Karlin. Method of invariant manifold for chemical kinetics. *Chemical Engineering Science*, 58(21):4751–4768, 2003.

[7] A.N. Gorban and I.V. Karlin. *Invariant manifolds for physical and chemical kinetics, Lect. Notes Phys. 660*. Springer Verlag, Berlin, Heidelberg, 2005.

[8] AN Gorban and O. Radulescu. Dynamic and static limitation in reaction networks, revisited . In David West Guy B. Marin and Gregory S. Yablonsky, editors, *Advances in Chemical Engineering - Mathematics in Chemical Kinetics and Engineering*, volume 34 of *Advances in Chemical Engineering*, pages 103–173. Elsevier, 2008.

[9] AN Gorban, O. Radulescu, and AY Zinovyev. Asymptotology of chemical reaction networks. *Chemical Engineering Science*, 65:2310–2324, 2010.

[10] S.A. Kauffman. *The origins of order: Self organization and selection in evolution*. Oxford University Press, USA, 1993.

[11] A.S. Matveev and A.V. Savkin. *Qualitative theory of hybrid dynamical systems*. Birkhauser, 2000.

[12] O. Radulescu, A.N. Gorban, A. Zinovyev, and A. Lilienbaum. Robust simplifications of multiscale biochemical networks. *BMC systems biology*, 2(1):86, 2008.

[13] M.A. Savageau and E.O. Voit. Recasting nonlinear differential equations as S-systems: a canonical nonlinear form. *Mathematical biosciences*, 87(1):83–115, 1987.

[14] R. Shorten, F. Wirth, O. Mason, K. Wulff, and C. King. Stability Criteria for Switched and Hybrid Systems. *SIAM Review*, 49(4):545–592, 2007.

[15] L. Tavernini. Differential automata and their discrete simulators. *Nonlin. Anal. Theory Methods Applic.*, 11(6):665–683, 1987.

Bioinformatic predictions and experimental validation of cis-regulatory modules in development: Application to cardiogenesis in *D.melanogaster*

Delphine POTIER[1,2], Stein AERTS[3], Carl HERRMANN[2] and Laurent PERRIN[1]

[1] IBDML – UMR6216 CNRS & Université de la Méditerranée, Marseille (France)
{potier,perrin}@ibdml.univ-mrs.fr
[2] TAGC – Inserm U928 & Université de la Méditerranée, Marseille (France)
herrmann@tagc.univ-mrs.fr.fr
[3] LCB, Dep. of Human Genetics, KU Leuven, Leuven (Belgium)

1 Introduction

Organogenesis and differentiation require the coordinated expression in time and space of different groups of genes. The accuracy of this process, governed by transcription factors (TFs) acting within a complex gene regulatory network, ensures the acquisition of specific organ shape and physiology. However, the logic of the cis-regulatory mechanisms is far from being understood so far. Bioinformatics approaches to predict cis-regulatory modules (CRM) from genomic sequences can greatly help to characterize new enhancers and the associated developmental regulatory network. Approaches based on combining expression data with comparative genomics are expected to allow predicting regions of DNA that regulate the expression of genes with greater accuracy. Previous approaches have been applied, based on predicting clusters of transcription factor binding sites, combined with phylogenetic footprinting [1,2], or using a statistical framework in order to decipher the most relevant combination of binding sites for expression-based gene clusters [3]. We present here a novel approach combining bioinformatics predictions of CRMs and experimental validations, which allowed us to identify CRMs from gene expression data.

2 Data and results

We focus our interest on the development of the cardiovascular system in Drosophila in order to investigate the regulatory logic of this process. During embryogenesis, cardiogenesis is mediated by a gene regulatory network (GRN) which includes conserved signaling pathways and transcription factors and leads to the formation of a linear cardiac tube, with antero-posterior polarity driven by the Hox genes. Then, during metamorphosis, the larval cardiac tube is remodeled to form the adult organ. We recently reported a precise temporal map of gene expression of adult heart formation through the analysis of the temporal dynamics of heart-specific gene expression profiling [4].

Starting from clusters of co-expressed genes during cardiac tube remodeling during metamorphosis, we applied a new method that uses a comprehensive library of position weight matrices, combined with phylogenetic conservation, to identify potential cis regulatory modules common to a cluster of co-expressed genes. Using this method, we have been able to predict several CRMs involving a particular class of TFs for one of the clusters, in which gene expression is induced at 42h after pupation. Potential binding sites are evolutionary conserved and overrepresented in the surrounding non-coding sequences of co-expressed genes with a high statistical significance. The class of TFs involved is likely to correspond to nuclear receptors, of which the drosophila homolog, *Dhr3*, is highly expressed during heart remodeling, on the onset of the induction of the cluster of genes. Besides this nuclear receptor, the predicted CRMs contain high confidence potential binding sites for MyoD like factors, which are specific of muscular tissues. We have performed in vivo validations, using transgenesis, using gateway cloning. The results show that the predicted CRMs reproduce the expected temporal expression pattern. Indeed, all six tested CRMs drive a transitory expression in different tissues from 42h to 96h after pupation. Our approach hence was successful in identifying CRMs regulating the temporal activation of the target genes, and our results suggests a modular architecture of the regulatory machinery, in which the temporal and spatial regulation are distinct.

We are performing further experimental validations, including mutagenesis of the predicted binding sites and transgenic assays in gain- and loss-of-function context for the predicted TF *Dhr3* to confirm the validity of the predicted CRMs.

References

1. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, et al. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. U.S.A 99: 757-762. doi:10.1073/pnas.231608898
2. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, et al. (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol 5: R61. doi:10.1186/gb-2004-5-9-r61
3. Philippakis AA, Busser BW, Gisselbrecht SS, He FS, Estrada B, et al. (2006) Expression-guided in silico evaluation of candidate cis regulatory codes for Drosophila muscle founder cells. PLoS Comput Biol 2: e53.
4. Zeitouni B, Sénatore S, Séverac D, Aknin C, Sémériva M, et al. (2007) Signalling Pathways Involved in Adult Heart Formation Revealed by Gene Expression Profiling in Drosophila. PLoS Genet 3: e174. doi:10.1371/journal.pgen.0030174

Parametric robustness in gene networks: reliable functioning with unreliable components

Ovidiu RADULESCU[1,2], Alexander N. GORBAN[3] and Andrei ZINOVYEV[4]

[1] DIMNP - UMR 5235 CNRS/UM1/UM2, Pl. E. Bataillon, Univ. of Montpellier 2, Bat. 24 CP 107, 34095 Montpellier Cedex 5, France
ovidiu.radulescu@univ-montp2.fr
[2] Team Symbiose, INRIA-IRISA, Campus de Beaulieu, 35042 Rennes, France
[3] University of Leicester, LE1 7RH Leicester, UK
[4] Institut Curie, U900 INSERM/Curie/Mines ParisTech, 26 rue d'Ulm, F75248 Paris, France

Keywords gene networks, robustness, error-correction

Robustness, defined as the capacity of a system to function reliably with unreliable components or to adapt to changing external conditions, represents a common feature of living systems. The fittest organisms are those that resist to diseases, to imperfections or damages of regulatory mechanisms, and that can function reliably in various conditions. There are many theories that describe, quantify and explain robustness. Waddington's canalisation [1] was formalised by Thom [2] as structural stability of attractors under perturbations. The canalization by attractors have been recently proven for Drosophila development [3]. The new field of systems biology places robustness in a central position among the living systems organizing principles, identifying redundancy, modularity and negative feedback as sources of robustness [4]. As noticed by von Dassow [6], systems biology models are robust with respect to variations of their parameters. Parametric robustness of models is also expressed by the strong anisotropy of sensitivity coefficients along directions in the parameter space (sloppy sensitivity). Robustness does not exclude fragility [4], as some of the model parameters could have a critical influence on the behavior of the system.

We discuss here system robustness with respect to randomness of the parameters. Our results can be applied to gene networks that function reliably with large variability in the strength of interactions between components. We formally define reliability as small variability of quantities defining network's functioning or output. We want to understand the general principles leading to robust functioning, but also to spot eventual fragility points that can be used to control the network.

Early insights into this problem can be found in the von Neumann's discussion of robust coupling schemes of automata [5]. von Neumann noticed the intrinsic relation between randomness and robustness. Quoting him "without randomness, situations may arise where errors tend to be amplified instead of cancelled out; for example it is possible that the machine remembers its mistakes, and thereafter perpetuates them". To cope with this, von Neumann introduces multiplexing and random perturbations in the design of robust automata.

We distinguish [8,9] three generic types of parametric robustness: simplex concentration, cube concentration and robust/fragile systems (systems with small number of critical parameters). The first two types can be related to the mathematical theory of concentration phenomena in high-dimensional spaces [7]. Model reduction techniques [11,10] can be used to identify critical processes and design rules leading to various robustness situations.

Simplex concentration and dominance effects are largely responsible for "sloppy sensitivity" phenomena, involving inequivalent contributions of elementary dynamical processes to the behavior of the system. Gene networks are multiscale systems, meaning that they involve wide ranges of protein abundances (from one to 10^4 per cell) and time scales of elementary dynamical processes, for instance biochemical reactions (from 10^{-3} to $10^4 s$). Contribution of these elementary dynamical processes to the behavior of the system is highly uneven. Thus, one process is dominating over many others and can be called critical [11]. Mathematically, system's dynamical properties depend on order statistics [9] (combinations of max or min over many parameters or parameter combinations). Order statistics have small variability even if parameter variation range is large, a phenomenon that is called simplex concentration.

Model reduction techniques for multi-scale network models extract the dominant sub-system and identify the critical parameters [11,10]. The model reduction algorithm contains pruning steps that eliminate dominated processes. These processes have little influence on the dynamics, which explains the overall sloppy

sensitivity of the model. As a result, a system with a small number of critical parameters is a paradigm for the robust/fragile concept.

Cube concentration produces reduced variability when many equivalent contributions are added together [9]. This phenomenon generalizes the law of large numbers. Properties showing cube concentration depend on many parameters of the dominant subsystem. An example of property having such behavior is the period of large oscillating networks [9].

We proposed a scenario to test various types of robustness [9]. In this scenario the variability of a given property (quantified by its log-variance) is computed for random variations of the parameters in two cases: i) all n parameters are changed independently with increasing individual log-variance, and ii) $r \leq n$ parameters are randomly chosen and then randomly changed with fixed log-variance for increasing values of the integer r. The two resulting plots (log-variance of the property as a function of the log-variance of the parameters in one case, and as a function of the number r of changed parameters in the second case) are discriminating for the three types of generic robustness. We have thus shown that for an oscillating signalling network the period of the oscillations follows cube concentration, the largest relaxation time follows simplex concentration, and the damping time of the oscillation amplitude is robust/fragile [9].

As a new development we present the application of this test to a large set of models from BioModels database for a large set of dynamical properties. We use a similar analysis, in the context of early development stages of Drosophila, to study the robustness of the cis-regulatory modules controlling the expression of even-skipped segmentation genes [12]. These studies illustrate the genericity of the mechanism.

Understanding robustness has fundamental importance as it can guide thinking about biological systems. It is important to known whether the control of a property of a system should be distributed (the case of properties with cube or simplex concentration) or localized on a well chosen target (the case of robust/fragile properties). Our studies also provide tools to identify the various types of robustness and the set of critical parameters which are important for practical applications. These tools complement more traditional sensitivity studies approaches. An even more important practical consequence of our results is the possibility to cope with parametric uncertainty of gene networks in a rational way. Indeed, determination of the dominant subsystems of a given multiscale network depends on the qualitative order relation and not on the precise values of the parameters. Determination of these order relations (qualitative comparison of interaction strengths by experimental techniques or by sequence analysis) allow simplification of the dynamics via model reduction tools and lead to identification of critical parameters that need to be measured more carefully.

References

[1] Waddington CH. The strategy of genes. London: Allen and Unwin; 1957.

[2] Thom R. Structural Stability and Morphogenesis. New York: Benjamin; 1975.

[3] Manu, Surkova S, Spirov AV, Gursky VV, Janssens H, Kim A-R, Radulescu O, Vanario-Alonso CE., Sharp DH., Samsonova M., Reinitz J. Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation. PLoS biology 2009;7(3):e1000049.

[4] Kitano H. Biological robustness. Nature Reviews. 2004;5:826–837.

[5] von Neumann J. Probabilistic logics and the synthesis of reliable organisms from unreliable components. In: J. von Neumann Collected works vol.5. Oxford: Pergamon Press; 1963. .

[6] von Dassow G, Meir E, Munro EM, Odell GM. The Segment Polarity Network is a Robust Developmental Module. Nature. 2000;406:188–192.

[7] Gromov M. Metric structures for Riemannian and non-Riemannian spaces, Progr.Math. 152. Boston: Birkhauser; 1999.

[8] Radulescu O, Gorban A, Vakulenko S, Zinovyev A. Hierarchies and modules in complex biological systems. Proceedings of ECCS'06. 2006.

[9] Gorban AN, Radulescu O. Dynamical robustness of biological networks with hierarchical distribution of time scales. *IET Syst. Biol*, 1(4):238–246, 2007.

[10] Gorban AN, Radulescu O, Zinovyev A. Asymptotology of chemical reaction networks. *Chemical Engineering Science*, 65:2310–2324, 2010.

[11] Radulescu O, Gorban AN, Zinovyev A, Lilienbaum A. Robust simplifications of multiscale biochemical networks. *BMC systems biology*, 2(1):86, 2008.

[12] Janssens H, Hou S, Jaeger J, Kim A-R, Myasnikova E, Sharp D, Reinitz J. Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. Nature genetics 2006;38(10):1159-65.

Structural-alphabet motifs in protein loop structures: from structure to function

Leslie REGAD[1], Juliette MARTIN[2] and Anne-Claude CAMPROUX[1]

[1] Molécules Thérapeutiques in silico (MTi), Université Paris Diderot - Paris 7, UMR-S973 Inserm, Batiment Lamarck, 5e étage, 35 rue Hélène Brion, 75205 cedex 13, Paris, France
{leslie.regad, anne-claude.camproux}@univ-paris-diderot.fr

[2] Université de Lyon, Lyon, France ; Université Lyon 1; IFR 128; CNRS, UMR 5086 ; IBCP, Institut de Biologie et Chimie des Protéines, 7 passage du Vercors, Lyon, F-69367, France
juliette.martin@ibcp.fr

Abstract *Structural genomics efforts lead to the determination of new protein structures that often lack sequence and fold similarity to known proteins. Sequence and structure-based methods may not be sufficient to predict the molecular function of these proteins. For such cases, the identification of functional motifs gives useful clues for deducing the protein function.*
We describe a new statistical method dedicated to the extraction of motifs of interest in protein loops. This method is based on the structural alphabet HMM-SA and the statistic over-representation. Thanks to HMM-SA protein structures is encoded into sequences of structural letters allowing the application of algorithms developed for sequence analysis such as the notion of pattern/word exceptionality. Thus, as in DNA sequences, the statistic over-representation related to SCOP superfamilies is used to extract structural motifs of interest in protein loops. Our analyses of biological annotations suggest that some structural motifs strongly over-represented in a SCOP superfamily are involved in the protein function, such as calcium- or nucleotide-binding site. Motifs detected by this approach could be used for the annotation of uncharacterized proteins.

Keywords structural-alphabet motifs, functional motifs, protein loops, statistic over-representation

1 Introduction

The prediction of protein function is a very important challenge. For many proteins, the search of homologous proteins with known function provides no straight answer. In such cases, the prediction of functional sites can give useful clues for deducing the protein function. Two types of methods have been developed for binding site prediction. On the one hand, some methods exploit the conservation of motifs associated to binding sites, which is effective if binding sites present strong amino-acid conservation [1,2]. On the other hand, some methods exploit the tri-dimensional (3D) structure of binding sites [3,4]. Most of these methods need for the learning functional motifs the knowledge of the position of functional site, and the computation of structural alignment or geometric descriptors.

In this paper, we present an alternative strategy for functional motif identification, based on a structural alphabet and statistics to detect 3D-motifs with exceptionnal frequency. We focus on the functional motifs from protein loops and used the structural alphabet HMM-SA [5]. It is a collection of 27 structural prototypes of four residues called structural letters, allowing the simplification of all protein structures into uni-dimensional structural-letter sequences. The use of word over-representation is motivated by the observation that functional sites in DNA are subject to selection pressure, which is expected to result in uncommon (high or low) frequency [6,7]. In a previous study, we have shown that HMM-SA, used in conjonction with pattern exceptionnality in the structural-letter sequences, is an effective tool for the mining of protein loops [8]. Here, we investigate the link between structural words (series of 4 consecutive structural letter sequences) and protein function, by looking for words specific to superfamilies defined by SCOP [9]. The role of these motifs in the protein function is then analyzed using the biological annotation Swiss-Prot.

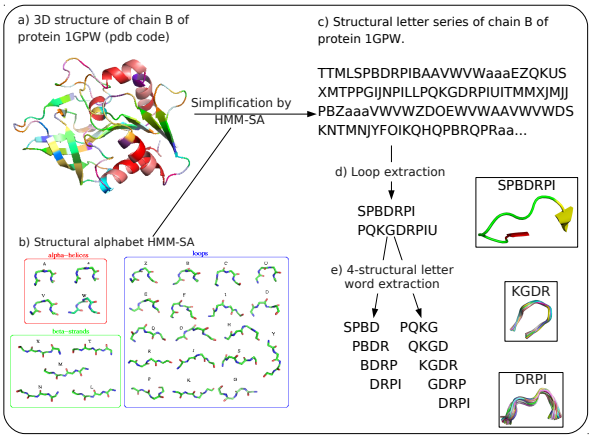


Fig. 1. Protocol of structural word extraction. a) the 3D structure of a structure is used as input, b) the Cα coordinates are simplified using HMM-SA, c) the result of the simplification is a sequence of structural letters, d) loops are extracted from the structural-letter sequences using regular expressions of structural letters, e) loops are systematically decomposed into overlapping words of four consecutive structural letters.

2 Material & Methods

Data set

An initial list of 8 119 protein structures was extracted from the PDB of May 2008 using the software PISCES [10] with the following criteria: obtained by X-ray diffraction, resolution better than 2.5 Å, longer than 30 residues, less than 50% sequence identity between any pairs. We restricted the list to the 5 429 structures classified in SCOP [9]. As it is assumed that proteins grouped in the same SCOP superfamily exhibit structural and functional similarities, this level was chosen for our analysis. To allow statistical analysis, we further restricted the list to the proteins classified in superfamilies with at least two members in the data set, resulting in 4 911 proteins from 1 493 superfamilies. On average, we found 7.90 proteins ($\pm$13.78) by superfamily.

Extraction of structural words

We described the 3D conformations of protein loops by structural words with the same protocol as in [8] and summarized in Fig. 1. It is based on the structural alphabet HMM-SA [5], a set of 27 structural prototypes of four residues, called structural letters, established using hidden Markov models.

Thanks to HMM-SA, a structure of n residues is encoded into a sequence of $(n-3)$ structural letters, where each structural letter describes the conformation of 4-residues.

The present study is focused on protein loops. We then discarded regular secondary structures from the structural letter sequences, based on the fact that some structural letters are specific to regular secondary structures [5,11]. The structural letter sequences of protein loops were further split into overlapping structural words of four letters (i. e. describing the conformation of seven residues).

We extracted a total number of 25 304 different structural words describing the conformation of 238 158 seven-residue fragments. As structural words with very low frequency could be linked to structural flexibility and regions with uncertain coordinates [8], we did not consider structural words seen less than 5 times. This resulted in a set of 11 294 words, accounting for 224 148 seven residue fragments, and seen, on average, 19.85 times ($\pm$31.69).

Computation of structural word statistics

The description of protein structures as sequences of structural letters allowed the application of algorithms developed for sequence analysis such as the notion of pattern/word exceptionality. The exceptionality of a word denotes its over- or under-representation in a data set. We used the SPatt software [12] to compute the exact statistics of structural word in sets of short sequences. It is achieved by comparing its real frequency in the data set and the frequency that would be expected under a background model defined as a first order Markov chain estimated on the set of protein loops. The over-representation score of a word w is given by:

$$Lp(w) = -\log_{10}[P(N^{exp}(w) > N^{obs}(w))]$$

where $N^{obs}(w)$ and $N^{theo}(w)$ denote respectively the frequency of w that is observed in the data set and expected under the background model, and P the probability of the event. A Lp score equal to 3 means that the pattern is over-represented with a p-value of 10^{-3}. To define the type of a word (over, under-represented or not significant), its Lp score is compared to a threshold. This significance threshold was defined by taking into account the multiple testing and was set to 5.97 using Bonferroni correction. Thus, a word with an over-represented score higher (resp. smaller) than 5.97 (resp. -5.97) is over-represented (resp. under-represented).

In order to investigate the link between structural words and function, the over-representation was assessed separately in each SCOP superfamily. In consequence, for each structural word, we have two criteria:

- L_{pmax}: the maximal L_p score among all superfamilies,

– nb_{sf*}: the number of superfamilies where a given word is significantly over-represented.

For the sake of comparison, we also computed these indicators in randomized data sets, obtained by randomly reassigning loops among SCOP superfamilies.

Biological annotations

Swiss-Prot is a curated sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, domain structure, post-translational modifications, variants, etc...), a minimal level of redundancy and high level of integration with other databases [13].

We chose Swiss-Prot database because it is a manually curated database and shows the lowest annotation error levels [14].

To map our structural words with Swiss-Prot annotations, we used the PDB/UniProt Mapping database [15], which consists of different files allowing the correspondence between PDB and UniProt codes, and PDB and Uni-Prot sequence numbering. Among the 4 911 protein structures, only 1 487 are included in the PDB/Uniprot Mapping database. The confrontation of structural words with biological annotations is thus inherently limited to a restricted data set.

We analyzed the correspondence between structural words and functional annotations available in Swiss-Prot database by counting the number of fragment of a word associated to an annotation in the data set.

External tools for prediction of protein features

Software REP [16] wad used to predict Repeat regions from protein sequences. It is an iterative homology-based Repeat finding method.

Software SitePredict [4] was used to predict nucleotide and calcium-binding sites. SitePredict is a machine learning-based method based on diverse residue-based properties including spatial clustering of residue types and evolutionary conservation. Only residues with a score higher than 0.5 are considered as residues involved in binding site.

3 Results & Discussion

Our goal is to systematically elicit structural motifs of interest extracted from protein loops.

3.1 Extraction of motifs of interest in loops using over-representation in SCOP superfamilies

We computed the over-represention of the 11 294 words in each superfamily. Fig. 2 presents these two statistic criteria for each word: Lp_{max}, the highest over-representation score over SCOP superfamilies and nb_{sf*} and the number of superfamilies where the word is over-represented. For example, the structural word `GSUS` is seen 169 times in 59 SCOP superfamilies with an Lp_{max} equal to 140 corresponding to its over-represented score in the superfamily "Pentapeptide repeat-like" (SCOP id 141571), meaning that it is strongly associated to this superfamily. It is also over-represented in the superfamilies "L domain-like" (SCOP id 52058) and "RNI-like" with over-representation scores of 40 and 7. Its nb_{sf*} is, thus, equal to 3. These two superfamilies have in common the property "contain amino-acid repeats". The consideration of nb_{sf*} thus permits to take into account the fact that some superfamilies could share a same over-represented structural motifs. In their study, Tendulkar et al, used a frequency parameter to define functional motifs: a cluster of fragment is functionally relevant if at least 70% of its fragments are extracted from proteins belonging to the same SCOP superfamily [17]. Using this criterion, the structural word `GSUS` is not functionally relevant, since only 27% and 15% of its fragments belong to the superfamilies "Pentapeptide repeat-like" and "L domain-like", respectively. Our indicators provide a more detailed analysis, since statistical over-representation takes into account the amount of data available in each superfamily.

data-set	Word nb[1]	Lp_{max}	nb_{sf*}
SCOP	11 294	4.3 (5.6)	0.2 (0.7)
SCOP random	11 294	2.5 (0.9)	0.006 (0.4)
Ubiquitous words	23	26 (14)	10.33 (5.5)
sf-specific words	24	89 (47)	1.4 (0.4)

Tab. 1. Average statistic parameters for different word sets. 1: number of words. In brackets is indicated the standard deviation.

We then checked the global significance of our results by comparing word statistics obtained using the actual SCOP classification and a randomized SCOP classification. We observe that Lp_{max} and nb_{sf*} are significantly higher in SCOP than random SCOP (cf. Tab. 1). This indicates that word over-representation in SCOP superfamilies is not random.

From Fig. 2, we make the distinction between three classes of structural words:

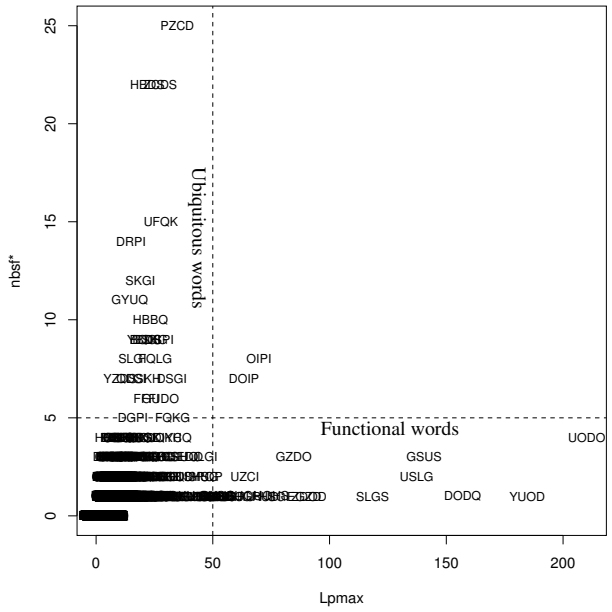


Fig. 2. Plot of the SCOP superfamily criteria for the 11 294 words. X-axis corresponds to Lp_{max} values measuring the tendency of words to be specific to one particular superfamily. Y-axis corresponds to nb_{sf*} values defined the number of superfamilies where words are over-represented.

- words being moderately over-represented in a large number of superfamilies, called ubiquitous words,
- words being highly over-represented in a small number of superfamilies, called sf-specific words,
- all other words are considered as uninformative, since that are neither sf-specific, nor ubiquitous.

The over-representation of a structural word indicates a surprising high frequency in a superfamily with respect to a background model of protein loops. A structural word that is over-represented in a superfamily was presumably selected during evolution because it has an important role for protein structure or function. To analyze this hypothesis, we studied in more details the most significant ubiquitous and sf-specific words. To perform that, we compared ubiquitous word to known recurrent structural motifs and investigated the link between sf-specific words and the functional site of proteins using Swiss-Prot annotations.

3.2 Examples of ubiquitous words

In this section we analyzed the 23 most ubiquitous words, that means structural words over-represented in a lot of superfamilyes ($nb_{sf*} \geq 5$). This ubiquitous word set presents a nb_{sf*} and Lp_{max} higher than on all words (cf. Tab. 1), and they are found across superfamilies with different folds (cf. Fig. 3). This Fig. 3 presents an illustration of properties of structural word PZCD, which is over-reprensented in 25 superfamilies.

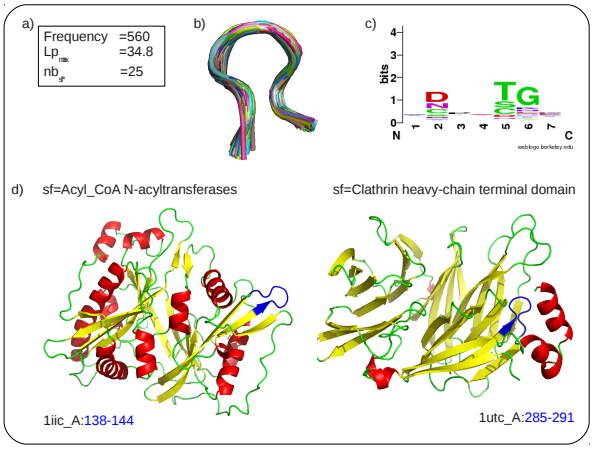


Fig. 3. Ubiquitous word PZCD. a) geometry of several PZCD-fragments optimally superposed. b) amino-acid conservation of the word generated by WebLogo [18]. c) word statistics. d) example of structures, represented using Pymol [19], containing the structural word PZCD

Some recurrent structural motifs in loops are well characterized and described in the literature, such as β-turns [20], α-turns [21] and γ-turns [22]. We carried out a comparison between our ubiquitous words and usual turn motifs. Following the definition of [20], turns were defined as tetrapeptides with a $C_i^\alpha - C_{i+3}^\alpha$ distance lower than 7 Å , with the two central residues $(i+1, i+2)$ in non-helical state, using the software ExtractTurn [23]. We then computed, for each ubiquitous word, the number of fragments containing a turn. Some words are not associated with turns: for example, SKGI and HBBQ have less than 7% of their fragments containing turns. On the contrary, other words are clearly associated to turns: all PZCD-fragments and 99% of FQKG-fragments contain a $\beta - turn$ in position 2-5 and 1-4, respectively. It is interesting to note that 1% of FQKG fragments fail the $\beta - turn$ assignment because they do not verify the threshold for the distance between $C_1^\alpha - C_4^\alpha$, which varies between 7.03 and 9.83 Å. In their paper, Fuchs et al. [24] defined "turn-like" conformations as corresponding to a four-residue fragment with a $C_1^\alpha - C_4^\alpha$ distance around 7 Å. This suggests that structural motifs could complement the classical geometric criteria to provide an assignation of turn-like fragments without relying on a strict distance threshold.

3.3 Examples of sf-specific words

In this section, we analyzed the 24 most significant sf-specific words, that means structural words with a Lp_{max} higher than 50, and a nb_{sf*} strictly smaller than 5.

Sf-specific words are strongly over-represented in 16 different superfamilies. Some superfamilies (EF-hand, P-loop containing nucleotide triphosphate hydrolases, S-adenosyl-L-methionine-dependent methyltransferases) group proteins containing small ligand-binding sites. Other superfamilies, e.g., L domain-like (52058), Pentapeptide repeat-like (141571) contain amino-acid repetitions. The statistic criteria of these most significant sf-specific words are presented in Tab. 2. Their very high over-representation in this superfamily could indicate a conservation in proteins of this superfamily during the evolution for functional reasons.

We thus wanted to know whether sf-specific words actually correspond to functional sites of proteins. We analyzed their association with functional annotations available in Swiss-Prot database [13]. The results of this analysis are presented in Tab. 2.

We can note that ten structural words, indicated in italic in Tab. 2, are strongly over-represented in several superfamilies poorly or not at all associated to a Swiss-Prot annotation. Six other sf-specific words are associated to Disulfide or Repeat annotations. The seven other sf-specific words are strongly associated the a functional annotation: nucleotide-binding site, calcium-binding site or binding annotations. Moreover, we can observe that some words are overlapping. For example YUOD and UODO, ZDOD and DODQ or SUQH and UQHS.

We here focused on three words UQHS, YUOD, DODQ (presented in Fig. 4) and analyzed in further details their link with SCOP superfamilies.

3.3.1 UQHS is a part of REPEAT regions

Structural word UQHS is strongly over-represented in the superfamily "L domain-like" (SCOP id 52058) grouping proteins containing Reapeat regions, i.e. repetition of amino-acid regions.

We can see that structural word UQHS presents amino-acid preferences (cf. Fig. 4) in close agreement with this consensus sequence of Repeat LRR: LxxLxLxxNxL or LxxLxLxxCxxL [25].

This word has 55% of its fragments annotated by the Repeat LRR (Leucine-Rich Repeat) annotation. This annotation indicates repeated sequence motifs or domains. This result shows that structural word UQHS is a part of Repeat regions (cf. Fig. 4) and suggests that unannotated UQHS-fragments correspond also to a part of Repeat regions.

To valide this hypothesis, we predicted repeat LRR, using software REP [16], in protein 1dce_A that contain two unannotated UQHS-fragments: 1dce_A:470-476 (amino-acid sequence=LSHNRLR) and

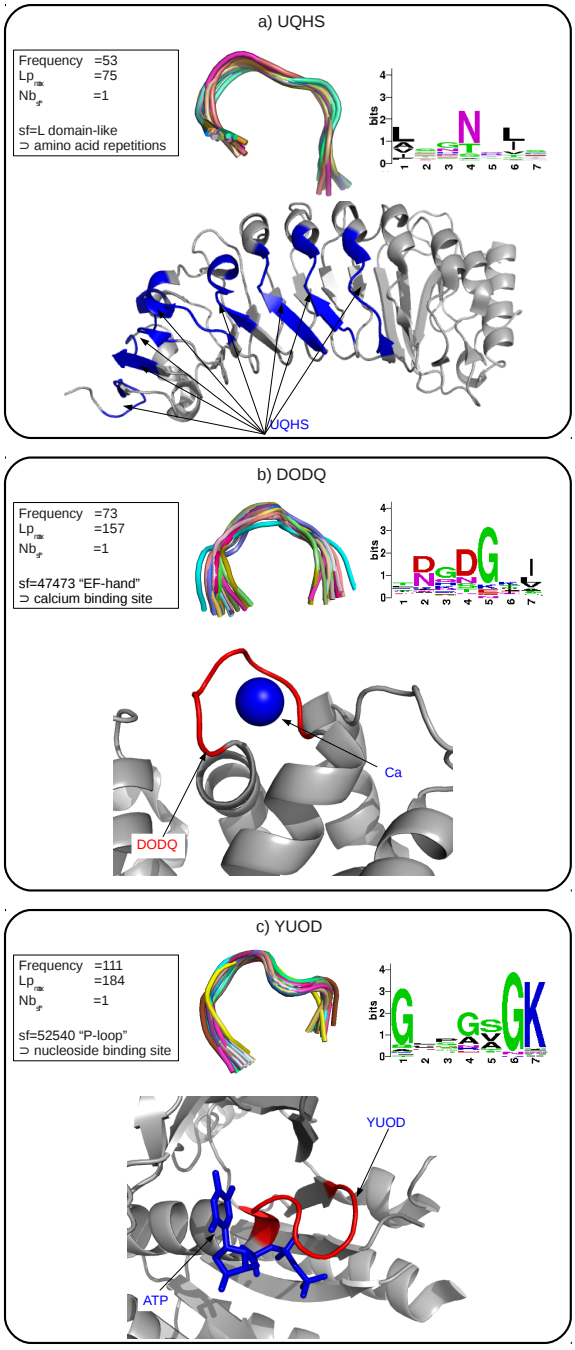


Fig. 4. Illustration of the three functional words : UQHS (a), DODQ (b) and YUOD (c). For each word, we provide: the statistics (word occurrence, Lp_{max}, nb_{sf*}), the superfamily where the word is the most over-represented, the superimposition of fragments associated to this word, the amino-acid conservation of using a Logo [18], and an example of the word in a protein structure. Motifs and protein structures are represented using Pymol [19].

Word	Lp_{max}	nb_{sf*}	annot	match/total **
UQHS	75.07	1	**Repeat**	12/22 (52)
SUQH	63.42	1	**Repeat**	11/26 (70)
QHSG	51.75	1	**Repeat**	4/10 (37)
HSGI	76.26	1	**Repeat**	5/12 (63)
QXUS	52.05	1	Repeat	1/15 (43)
ZSGI	52.22	1	Repeat	7/36 (99)
GSUS	140.49	3	Repeat	6/38 (169)
GZDO	84.72	3	Repeat	1/35 (115)
DODQ	157.01	1	**CA_BIND**	15/23 (73)
ZDOD	91.27	1	**CA_BIND**	11/16 (48)
YUOD	184.67	1	**NP_BIND**	39/41 (111)
UODO	210.14	4	**NP_BIND**	49/60 (142)
OEIJ	53.84	1	**NP_BIND**	6/7 (33)
EIJU	51.68	1	**NP_BIND**	7/15 (48)
USLG	137.35	2	NP_BIND	2/22 (121)
UZCI	63.70	2	NP_BIND	1/13 (99)
RUDO	55.55	1	**Binding**	5/10 (27)
URNH	54.95	1	**Disulfide**	7/14 (43)
RNHB	51.33	1	**Disulfide**	9/20 (59)
UGRU	60.07	1	Mutagen	1/12 (37)
EGZD	51.68	1		(48)
GRUD	70.55	1		(33)
SLGS	118.45	1		(60)

Tab. 2. Results of the sf-specific word annotations. ** match and total denote respectively the number of fragments that is annotated and the total number of fragments. This comparison with Swiss-Prot annotation is restricted to the set of proteins that are common to our datase and Swiss-Prot database. The number between brackets describes the total number of fragments in our data set. Bold font indicates ration match/total higher than 50%. Italic font indicates ration match/total smaller than 50%. NP_BIND: Extent of a nucleotide phosphate binding region. CA_BIND: Extent of a calcium-binding region.

1dce_A:493-499 (amino-acid sequence=ASDNALE). We found two predicted Repeat LRR in protein 1dce_A at positions 484-507 and 529-553. Region 484-507 overlapps the second unannotated UQHS fragment (493-499). The first UQHS unannotated fragment is close in the sequence to the first predicted LRR-Repeat and has an amino-acid sequence (LSHNRLR) in agreement with the consensus sequence of LRR-repeats, suggesting that it is a LRR-repeat.

We can conclude that the structural motif UQHS corresponds to a part of the highly conserved region of LRR-repeat, that is not defined as protein functional site. However, proteins with LRR repeat have strong biological implications: they are involved in protein-protein interactions in plant and mammalian immune response [26]. A number of human diseases have been shown to be associated with mutation in the genes encoding LRR-proteins, principally in LRR domains [26].

3.3.2 Some sf-specific words are involved in binding sites

DODQ ***is a calcium-binding site.*** Structural word DODQ is over-represented in the superfamily "EF-hand" (SCOP id=47473) grouping proteins with EF-hand units, made of two helices connected with calcium-binding loop.

This word presents amino-acid preferences (cf. Fig. 4) in agreement with the sequence consensus of the calcium-binding motifs [DxDxDG] [27].

Structural word DODQ is strongly associated to the calcium-binding site annotation (cf. Tab. 2). This result shows that structural word DODQ is a calcium-binding site, and suggests that the 9 unannotated fragments are calcium-binding sites.

To validate this hypothesis, we predicted, using SitePredict, the calcium-binding sites in proteins containing unannotated DODQ-fragments. Six out of the nine unannotated DODQ-fragments contain residues predicted as involved in calcium-binding sites. Thus, among the 23 DODQ-fragments, 20 correspond to annotated or predicted calcium-binding sites.

Rigden et al. extracted structural motifs from calcium-binding proteins in order to analyze the structural diversity of these proteins [27]. From the six proteins common to our data-set and theirs, they extracted 13 calcium-binding sites. The structural alphabet analysis of these sites shows that all correspond to structural motifs DODQ, except one which corresponds to DODS.

These results allow to conclude that structural word DODQ is involved in calcium-binding sites.

YUOD ***is a nucleotide-binding site.*** YUOD is strongly over-represented in the superfamily "P-loop containing nucleoside triphosphate hydrolases" (SCOP id 52540) grouping proteins with nucleotide-binding site.

YUOD presents a clear amino-acid conservation (cf. Fig. 4) in agreement with the one of P-loops: [AG]XXXXGK[TS] [28].

The structural motif YUOD is very strongly assciated to the nucleotide binding site annotation (cf. Tab. 2): only 2 fragments (1ogo_X:45-51, 1lwj_A:354-360) are not annotated by this annotation.

To investigate the functional role of these unannotated fragments, we predicted the nucleoside-binding site of these two proteins using the software SitePredict. These two proteins (1ogo_X, 1lwj_A) do not contain predicted nucleotide-binding site, that can not confirm the functional role of these two fragments.

In their study, Via et al. used 3DLogo to extract structural motifs in P-loops [29]. They illustrated their

method on six proteins containing P-loops, by predicting the residues involved in ATP/GTP-binding sites. From this protein set, we extracted six structural words corresponding to their predicted binding sites. Five of them are encoded into YUOD and one into KUED, two words that are structurally close.

These results allow to conclude that structural word YUOD is involved in nucleotide-binding sites.

3.3.3 Limits of our method We have analyzed in details the 23 most significant sf-specific words, and 20 of them could be mapped to a SwissProt annotation. Using relaxed parameters ($nb_{sf*} < 5$ and $Lp_{max} \geq 10$), the number of sf-specific words can be raised to 565. Among these 565 sf-specific structural words, only 46 words have an association with an Swiss-Prot annotation.

This deceptively low number of sf-specific words with confirmed functional implication can be explained by two elements: (i) only 30% of the proteins used in our study are mapped to SwissProt annotations in the PDB/UniProt Mapping Database, and (ii) for a given protein, annotations reflect the state of our current knowledge and could thus be incomplete. This important bottleneck introduced by Swiss-Prot probably lowers the effective information content of our structural words. For example, structural word UGRU is seen 37 times in the SCOP data set and is strongly over-represented in the superfamily "S-adenosyl-L-methionine-dependent methyltransferases" (SCOP id=53335) (cf. Table Tab. 2). In the Swiss-Prot protein data-set (1487 proteins), it is seen only 12 times, indicating in a loss of 65% of UGRU-fragments in this validation step. Out of these 12 fragments only one fragment is annotated by the "Mutagen" annotation (cf. Table Tab. 2). However, the manual analysis of the functional annotations of the 29 UGRU-fragments seen in the superfamily "S-adenosyl-L-methionine-dependent methyltransferases" through the Swiss-Prot web interface (http://www.uniprot.org/uniprot/) reveals that 12 fragments are annotated by the binding site to S-adenosyl-L-methionine (SAH/SAM) ligand. Among the 17 unannotated UGRU-fragments, 8 are extracted from proteins co-crystallized with SAH/SAM, and the UGRU-fragments contain residues involved in the SAH/SAM-binding site. Thus, 69% of UGRU-fragments are involved in a SAH/SAM-binding site

This is an example where the automatic validation using the PDB/UniProt Mapping Database clearly underestimate the real functional implication of a structural word.

4 Conclusion & Perspectives

We present a method allowing the extraction of tridimensional motifs from loops important for protein structure or function. This method is based on the simplification of loop structures using structural alphabet HMM-SA and the over-representation of motifs in a set of proteins with similar function, provided by SCOP superfamily classification.

The analysis of statistical over-representation of motifs in SCOP superfamilies allowed distinguishing two interesting classes of motifs: ubiquitous, i.e. words over-represented in lot of superfamilies, and sf-specific motifs, i.e. words over-represented in several superfamilies.

The comparison between ubiquitous words and small known 3D motifs showed that some of ubiquitous words contain β-turns motifs. They are specific to some superfamilies that shows the importance of these motifs for protein folding or function.

The analysis of the functional annotations of sf-specific motifs, provided by Swiss-Prot database, showed that some motifs are included in Repeat regions (=amino-acid repetitions). Some others are identified as involved in functional sites such as binding sites of small ligands (calcium, nucleotide, SAH/SAM).

Thus, this study showed that, as in DNA, over-representation is an effective tool for the extraction of motifs of interest involved in protein structure or function.

In this study, we showed that the identification of these sf-specific motifs in proteins suggests some annotations that are not containing in Swiss-Prot database. Thus, sf-specific words could be used to improve the functional annotation of proteins and add annotations in Swiss-Prot database. Moreover, the identification of these functional words in structural-letter sequences corresponding to the structure of uncharacterized proteins is useful for the prediction of functional sites in these proteins.

References

[1] C. Andreini, I. Bertini, and A. Rosato. A hint to search for metalloproteins in gene banks. *Bioinformatics*, 20(9):1373–1380, 2004.

[2] N. Shu, T. Zhou, and S. Hovmoller. Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics*, 24(6):775–782, 2008.

[3] B. J. Polacco and P. C. Babbitt. Automated discovery of 3D motifs for protein function annotation. *Bioinformatics*, 22:723–730, 2006.

[4] A.J. Bordner. Predicting small ligand binding sites in proteins using backbone structure. *Bioinformatics*, 24(24):2865–2871, 2008.

[5] A. C. Camproux, R. Gautier, and T. Tufféry. A hidden Markov model derivated structural alphabet for proteins. *J Mol Biol*, 339:561–605, 2004.

[6] M. Y. Leung, G. M. Marsh, and T. P. Speed. Over- and underrepresentation of short DNA words in herpesvirus genomes. *J Comput Biol*, 3:345–360, 1997.

[7] E. Rocha, A. Viari, and A. Danchin. Oligonucleotide bias in bacillus subtilis: general trends and taxonomic comparisons. *Nucl Acids Res*, 26:2971–2980, 1998.

[8] L. Regad, J. Martin, G. Nuel, and A. C. Camproux. Mining protein loops using a structural alphabet and statistical exceptionality. *BMC Bioinformatics*, 11:75, 2010.

[9] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247:536–540, 1995.

[10] G. Wang and R.L. Dunbrack. Pisces: a protein sequence culling server. *Bioinformatics*, 19:1589–1591, 2003.

[11] L. Regad, J. Martin, and A. C. Camproux. Identification of non random motifs in loops using a structural alphabet. In *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational*, pages 92–100, Toronto, September 2006.

[12] G. Nuel, L. Regad, J. Martin, and A. C. Camproux. Exact distribution of pattern in a set of random sequences generated by a Markov source: application to biological data. *Algo Mol Biol*, 5:15, 2010.

[13] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. The universal protein resource (UniProt). *Nucl Acids Res*, 33:154–159, 2005.

[14] A.M. Schnoes, S.D. Brown, I. Dodevski, and P.C. Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, 5(12):–, 2009.

[15] A.C.R. Martin. Mapping pdb chains to uniprotkb entries. *Bioinformatics*, 21(23):4297–4301, 2005.

[16] M.A. Andrade, C.P. Ponting, T.J. Gibson, and P. Bork. Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol.*, 298(3):521–537, 2000.

[17] A. V. Tendulkar, A. A. Joshi, M. A. Sohoni, and P. P. Wangikar. Clustering of protein structural fragments reveals modular building block approach of nature. *J Mol Biol*, 338:611–629, 2004.

[18] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. WebLogo: A sequence logo generator. *Genome Res*, 14:1188–1190, 2004.

[19] W. L. DeLano. The pymol molecular graphics system (2002) on world wide web http://www.pymol.org.

[20] P. N. Lewis, F. A. Momany, and H. A. Scheraga. Chain reversals in proteins. *Biochim Biophys. Acta*, 303(2):211–229, 1973.

[21] V. Pavone, G. Gaeta, A. Lombardi, F. Nastri, O. Maglio, C. Isernia, and M. Saviano. Discovering protein secondary structures: classification and description of isolated α-turns. *Biopolymers*, 38:705–721, 1996.

[22] E. J. Milner-White, B. M. Ross, R. Ismail, K. Belhadj-Mostefa, and R. Poet. One type of gamma-turn, rather than the other gives rise to chain reversal in proteins. *J Mol Biol*, 204:777–782, 1988.

[23] P. Fuchs, J. F. Alix, and J. P. Alain. High accuracy prediction of beta-turns and their types using propensities and multiple alignments. *Proteins*, 59:828–839, 2005.

[24] P.F.J. Fuchs, A.M.J.J. Bonvin, B. Bochicchio, A. Pepe, A.J.P. Alix, and A.M. Tamburro. Kinetics and thermodynamics of type viii beta-turn formation: a cd, nmr, and microsecond explicit molecular dynamics study of the gdnp tetrapeptide. *Biophys. J.*, 90(8):2745–2759, 2006.

[25] A.V. Kajava. Structural diversity of leucine-rich repeat proteins. *J. Mol. Biol.*, 277(3):519–527, 1998.

[26] N. Matsushima, N. Tachi, Y. Kuroki, P. Enkhbayar, M. Osaki, M. Kamiya, and R.H. Kretsinger. Structural analysis of leucine-rich-repeat variants in proteins associated with human diseases. *Cell Mol. Life Sci.*, 62(23):2771–2791, 2005.

[27] D. J. Rigden and M. Y. Galperin. The DxDxDG motif for calcium binding: multiple structural contexts and implications for evolution. *J Mol Biol*, 343:971–984, 2004.

[28] M. Saraste, P. R. Sibbald, and A. Wittinghofer. The P-loop: a common motif in ATP- and GTP-binding proteins. *Trends Biochem Sci*, 15:430–434, 1990.

[29] A. Via, D. Peluso, P. F. Gherardini, E. de Rinaldis, T. Colombo, G. Ausiello, and M. Helmer-Citterich. 3dLOGO: a web server for the identification, analysis and use of conserved protein substrucutres. *Nucl Acids Res*, 35:W416–9, 2007.

Protein sequences classification by means of feature extraction with substitution matrices

Rabie SAIDI[1,2], Mondher MADDOURI[2] and Engelbert MEPHU NGUIFO[1]

[1] LIMOS, UMR 6158 CNRS, F-63173 Aubière, France
{saidi, mephu}@isima.fr
[2] URPAH, Université de Tunis El Manar – FST, Campus Universitaire le Belvédère 1960, Tunis, Tunisie
mondher.maddouri@fst.rnu.tn

Abstract *This work was recently published in BMC Bioinformatics [1] and a preliminary version was reported in [2]. It deals with the preprocessing of protein sequences for supervised classification. Motif extraction is one way to address that task. It has been largely used to encode biological sequences into feature vectors to enable using known machine-learning classifiers which require this format. However, designing a suitable feature is not a trivial task. For this purpose, we propose a novel encoding method that uses amino-acid substitution matrices to define similarity between motifs during the extraction step. We carried out various experiments to compare with existing approaches. The outcomes confirm the efficiency of our encoding method to represent protein sequences in classification tasks.*

1 Proposed Method

1.1 Overview

The DDSM (Discriminative Descriptors with Substitution Matrix) encoding method is composed of three parts [1]. First, we extract discriminative substrings using an adaptation of the Karp, Miller and Rosenberg (KMR) algorithm [3]. A motif is considered to be discriminative between a family *F* and other families if it appears in *F* significantly more than it does in the other families. Second, we keep only one motif for each cluster of substitutable motifs of the same length. Third, we construct an object-property table where objects are protein sequences and properties are motifs. We denote by 1 the presence of a motif or of one of its substitutes and 0 otherwise.

The second part can be also divided into two phases: (i) identifying clusters' main motifs and (ii) filtering.

i - The main motif of a cluster is the one that is the most likely to mutate to another in its cluster. To identify all the main motifs, we sort $\mathcal{M}$ in a descending order by motif lengths, and then by P_m (probability of mutation to another motif [1]). For each motif *M'* of $\mathcal{M}$, we look for the motif *M* which can substitute *M'* and that has the highest P_m. The clustering is based on the computing of the substitution probability between motifs [1]. We can find a motif which belongs to more than one cluster. In this case, it must be the main motif of one of them (see table 1 and figure 1). We draw attention that all measures concerning the substitution among motifs are derived from substitution matrices.

ii - The filtering consists of keeping only the main motifs and removing all the other substitutable ones. The result is a smaller set of motifs which can represent the same information as the initial set.

1.2 Illustrative Example

Given a Blosum62 substitution matrix and the following set of motifs (table 1) sorted by their lengths and P_m, we assign each motif to a cluster represented by its main motif. We get 5 clusters illustrated by the diagram shown in figure 1. This figure illustrates the set of clusters and main motifs obtained from the data of table 1 after application of our algorithm.

$\mathcal{M}$	LLK	IMK	VMK	GGP	RI	RV	RF	RA	PP
P_m	0.89	0.87	0.86	0	0.75	0.72	0.72	0.5	0
Main motif	LLK	LLK	LLK	GGP	RI	RI	RI	RV	PP

Tab. 1. Motif clustering

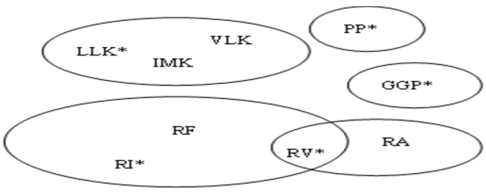


Fig. 1. Motif clustering. Kept motifs are : LLK, GGP, RI, RV and PP.

2 Experiments and Results

We tried to conduct our experiments on various kinds of datasets. These datasets differ from one another in terms of size, number of class, class distribution, complexity and sequence identity percentage. In this paper we present only two already published datasets (DS1 and DS2 in table 2). This allowed us to carry out a comparison with several related works. DS1 [4] contains seven classes that represent seven categories of quaternary protein structure with a sequence identity of 25% extracted from Swiss-prot. The problem here lies in recognizing the 4D structure category from the primary structure. DS2 consists of 277 domains: 70 all-α domains, 61 all-β domains, 81 α/β domains, and 65 α+β domains from SCOP. This challenging dataset was constructed by Zhou [5] and has been extensively used to address structural class prediction [5,6].

Dataset	Identity percentage	Family/class	Size	Total
DS1	25 %	Monomer	208	717
		Homodimer	335	
		Homotrimer	40	
		Homotetramer	95	
		Homopentamer	11	
		Homohexamer	23	
		Homooctamer	5	
DS2	84 %	All-α domain	70	277
		All-β domain	61	
		α / β domain	81	
		α + β domain	65	

Tab. 2. Experimental data

We combine our encoding method with several known classifiers (decision tree C4.5, naïve bayes NB, support vector machines SVM and nearest neighbour NN). We compare with several approaches reported in [4,5,6] and we report the results (accuracy rates) of the experiments on DS1 and DS2 in figure 2 and 3.

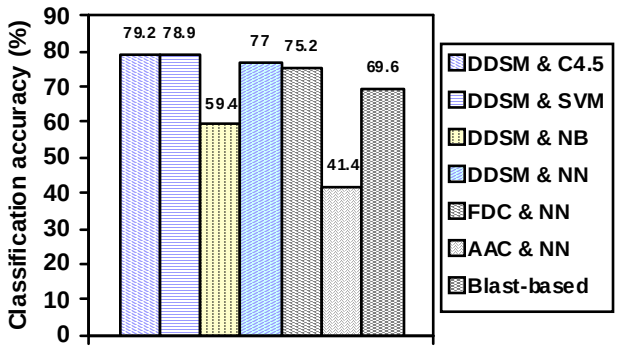


Fig. 2. Comparison with results reported in (Yu et al., 2006) for DS1. Details in [1].

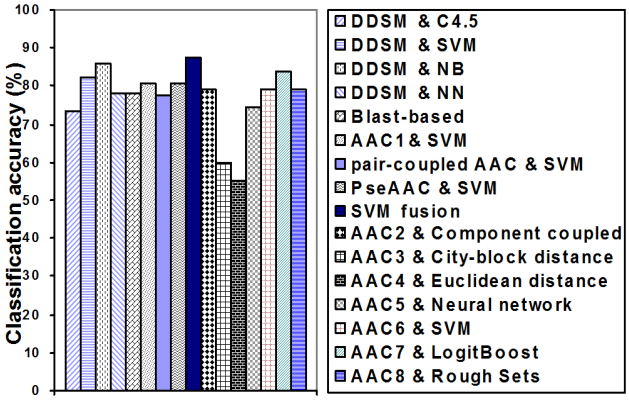


Fig. 3. Comparison with results reported in (Chen et al., 2006) and (Zhou, 1998) for DS2. Details in [1].

In both figure 2 and 3, we can notice that DDSM (first four histograms) allows reaching high accuracies. However, in related works, authors perform a fine-tuning to look for the classifier parameter values allowing to get the best results, whereas we just use the default parameter values of both our encoding method and the classifiers.

Acknowledgements

This work was partially supported by the French-Tunisian project CMCU-Utique 05G1412 and LifeGrid PREFON_META project.

References

[1] R. Saidi, M. Maddouri and E. Mephu Nguifo, Protein sequences classification by means of feature extraction with substitution matrices, *BMC Bioinformatics*, **11**:175, 2010.

[2] R. Saidi, M. Maddouri and E. Mephu Nguifo, 2007, "Evaluation of Biological Sequences Encoding for Supervised Classification, JOBIM, Poster, Marseille, 9-12 Juillet, 409-410 2007.

[3] R. Karp, R.E. Miller, A.L. Rosenberg, Rapid Identification of Repeated Patterns in Strings, Trees and Arrays, *4th Symposium of Theory of Computing*:125-136, 1972.

[4] X. Yu, C. Wang, Y. Li, Classification of protein quaternary structure by functional domain composition. *BMC Bioinformatics*, **7**:187-192, 2006.

[5] G.P. Zhou, An intriguing controversy over protein structural class prediction. *J. Protein Chem*, **17**: 729–738, 1998.

[6] C. Chen, X.B. Zhou, Y.X. Tian, X.Y. Zhou, P.X. Cai, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. *Anal Biochem*, **357**: 116–121, 2006.

Présentations courtes

Characterization of the Bcl-2 family using structure-aided HMM framework

Abdel AOUACHERIA, Valentine RECH DE LAVAL, Gilbert DELÉAGE and Christophe COMBET

IBCP, UMR5086, CNRS, Université Lyon 1, 7 passage du Vercors, 69367, Lyon, Cedex 07, France
{a.aouacheria, c.combet}@ibcp.fr

Keywords Apoptosis, Bcl-2 family, HMM profiles, indels, remote homology, database.

1 Introduction

The Bcl-2 family controls induction of apoptosis (programmed cell death) at the mitochondria via opposing functions of prosurvival and proapoptotic regulators [1]. At the level of primary structure, members of this family are currently classified based on the presence of one or more Bcl-2 Homology (BH) domains, which participates in the formation of homo- and hetero-dimers [2]. Antiapoptotic multidomain proteins ('Bcl-2-like' members) are considered to share sequence similarity in four BH domains (BH1–4). Proapoptotic proteins are divided into two subgroups: proapoptotic multidomain proteins ('Bax-like' members), which are assumed to contain only two or three BH domains (BH1-3), and proapoptotic BH3-only proteins (including Bid or Bim), which have sequence similarity only in the BH3 domain, a short amphipathic α-helix.

Recent phylogenomic [3], bioinformatics [4] and structural studies [5] highlighted the extent of sequence variation between Bcl-2 family proteins and raised several important points. First, while they appear to be *bona fide* family members, a number of multi-BH proteins lack the BH1 or BH4 domains or do not have any recognizable BH3 signature. Moreover, despite being classified as a BH3-only protein, Bid exhibits a 3D structure with a fold identical to that of Bcl-2 and Bax. This conserved Bcl-2 'core' fold is composed of a globular bundle of 5-7 amphipathic helices surrounding one central hydrophobic α-helix. However, except for Bid, most BH3-only proteins have unrelated predicted structures and some were assigned to the class of intrinsically unstructured proteins. Last, highly divergent viral homologues sharing the same helical fold as Bcl-2 but with virtually no recognizable sequence similarity have recently been reported. All these lines of evidence point to a need (i) to redefine the specific structural features and sequence signatures of the extended Bcl-2 family; (ii) to capture expert knowledge and integrate novel data into a dedicated database, such as the recently developed BCL2DB [6].

2 Results

We have developed a set of profiles specific for the various Bcl-2 family subgroups (homologous helix-bundled cellular and viral members, and related BH3-only and BH3 domain-containing proteins) using profile-based hidden Markov models (HMMs) combining sequence and structure information with the HMMER [7] package and the NPS@ web server [8]. These unique HMM profiles of conserved residues were compared to standard profiles (e.g. PFAM) for Bcl-2 family/BH3 domain recognition and used for database searches and pangenomic queries. We checked the results for sensitivity to include all presumed members of the family and verified that the HMM-detected sequences do not overlap with other known families. Our different models can be useful for improving the power of computational annotations (classification) and testing for potential membership in the family, including that of novel cellular and viral sequences with vanishingly low sequential similarity. As an example, we report the discovery of BCL-WAV (Acc#: D2Y5Q2), a divergent Bcl-2 homolog found in water-living anamniote vertebrates (fishes and anurans). BCL2DB will be expanded with an update system to automatically include these predicted members along with Bcl-2 family proteins with known (experimentally confirmed) functions.

By exploiting knowledge of the conserved block positions, we also analyzed insertion/deletion events (indels) occurring in the sequences of vertebrate Bcl-2 family proteins adopting the Bcl-2-like topology. Our data suggest that indels represent an important source of genetic and structural divergence between family members (paralogs) and species homologs (orthologs), likely to translate into functional diversity. These signatures were used as phylogenetic markers to propose a sequence of events leading up to the present-day repertoire of helix-bundled Bcl-2 family proteins.

3 Conclusions and future work

Computational redefinition of the Bcl-2 family and identification of distantly related members (e.g. viral homologues) represent major challenges to accelerate the functional understanding of massively available sequences issued from genomic and post-genomic efforts and to reveal remote evolutionary links. Analysis of the full set of known and novel protein sequences retrieved by the HMM profiles will be used to update BCL2DB and its classification system.

Acknowledgements

VRL is supported by a doctoral fellowship from La Ligue Contre le Cancer (Comité de Saône et Loire). This project is developed on the GIS-IBiSA PRABI bioinformatics platform.

References

[1] RJ. Youle and A. Strasser, The BCL-2 protein family: opposing activities that mediate cell death. Nat Rev Mol Cell Biol., 9:47-59, 2008.

[2] JM. Hardwick and RJ. Youle, SnapShot: BCL-2 proteins. Cell., 138(2):404, 404.e1, 2009.

[3] A. Aouacheria, F. Brunet and M. Gouy, Phylogenomics of life-or-death switches in multicellular animals: Bcl-2, BH3-Only, and BNip families of apoptotic regulators. Mol Biol Evol,. 22:2395-416, 2005.

[4] D. Lama and R. Sankararamakrishnan. Identification of core structural residues in the sequentially diverse and structurally homologous Bcl-2 family of proteins. Biochemistry,. 49:2574-84, 2010.

[5] M. Kvansakul, H. Yang, WD. Fairlie, PE. Czabotar, SF. Fischer, MA. Perugini, DC. Huang, PM. Colman. Vaccinia virus anti-apoptotic F1L is a novel Bcl-2-like domain-swapped dimer that binds a highly selective subset of BH3-containing death ligands. Cell Death Differ,. 15:1564-71, 2008.

[6] A. Aouacheria and SV. Blaineau. BCL2DB: moving 'helix-bundled' BCL-2 family members to their database. Apoptosis,. 14:923-5, 2009.

[7] Eddy SR, A new generation of homology search tools based on probabilistic inference. Genome Inform. 23, 205-211, 2009.

[8] C. Combet, C. Blanchet, C. Geourjon and G. Deléage. NPS@: Network Protein Sequence Analysis. Trends Biochem Sci., 25, 147-150, 2000.

Fast and accurate multiple sequence alignment of large and diversified sets of distant homologues

Matthieu BARBA, Olivier LESPINET and Bernard LABEDAN

Institut de Génétique et Microbiologie, UMR8621 CNRS,
Université Paris-Sud XI, Bâtiment 400, 91405 Orsay Cedex, France
{matthieu.barba, olivier.lespinet, bernard.labedan}@igmors.u-psud.fr

Abstract *Frali allows delivering an accurate and biologically relevant multiple sequence alignment (MSA) of large and heterogeneous families comprising remote homologues. First, an expert alignment of well-studied representatives of each subfamily is built semi-manually to define a seed alignment that represents the frame of the whole family. Then; the targeted addition of the rest of the parental sequences to this frame is processed after being sampled according to their degree of relatedness to their homologues prealigned in the frame. These new sequences are further clustered before aligning them to this frame using a hidden Markov model based profile-profile approach. This process allows keeping the accuracy gained at the step of building the seed alignment as checked both by benchmarking and by studying a family of distant homologous enzymes involved in various biological functions. Interestingly, this approach further allows a rapid update of a reference MSA as soon as new homologues appear.*

Keywords multiple alignment, remote homologues, HMM profile.

Aligner rapidement et exactement de grands jeux d'homologues distants

Résumé *Pour obtenir un alignement multiple exact et biologiquement valide de séquences homologues distantes appartenant à des grandes familles hétérogènes, une graine formée des représentants caractéristiques de chaque sous-famille est construite pour représenter l'architecture de la famille. Puis, le reste des séquences homologues à cette graine est ajouté progressivement de façon complètement automatisée par une approche profil-profil de modèles cachés de Markov. Cette approche permet de maintenir la qualité optimale de la-graine et (cerise sur le gâteau) de mettre à jour automatiquement à tout moment l'alignement de référence.*

Mots-clés *alignement multiple, homologues distants, profil HMM.*

1 Introduction

Many biologists consistently use completely automatic tools to generate multiple sequence alignment (MSA) without considering their potential flaws. In fact, although many algorithms are now available [1,2,3], constructing a MSA is not a trivial task [4]. Since defining homology is always a hypothesis, only empirical approaches are suitable. Hence, as already underlined [3], MSA are not plain data but models. Therefore, manual construction still remains more appropriate than automated one to get biologically relevant MSA [4], and if an automatic approach is used, a manual check is obligatory to improve the obtained output. It becomes increasingly difficult to meet these requirements as the number of potential homologues increases vertiginously.

Presently, families containing several thousands of homologues have become common, making it mandatory to use a limited number of automated tools, such as Muscle [5], while rendering difficult the required manual check of the output. Moreover, computing such alignments of large sets of sequences in a reasonable time implies a concomitant loss in correctness. Indeed, Kemena and Notredame [3] showed that the present MSA methods lose their accuracy when the number of sequences to multiply align is >100.

The challenge of building accurate MSA becomes even harder when dealing with distantly related homologous proteins. This often occurs in large and diversified families where subfamilies may be very distant from each other, their amino acid sequences sharing very low percentage of sequence identity as exemplified in the test case described below (§ 4.2).

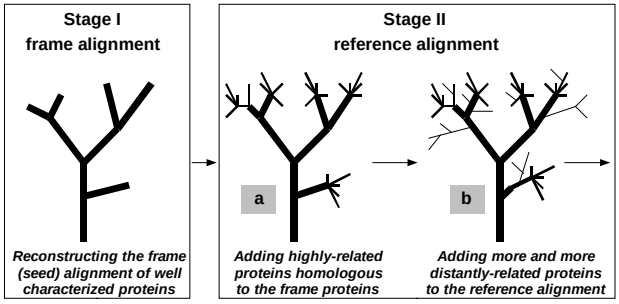


Fig. 1. Stepwise automated generation of an accurate reference alignment of a large and assorted family (stage II) using as a template a biologically relevant frame alignment built semi-manually (stage I).

To cope with these technical limitations, we propose a 2-stage approach based on (i) the construction of a high quality seed alignment, using a small, selected, set of sequences, and (ii) the progressive and targeted addition of the rest of the homologues to this seed (Fig. 1). By definition, a seed alignment would be optimal where each site corresponds to homologous positions, i.e. if each column contains the amino acids believed to have evolved from a common ancestor only through character substitutions [6,7]. For this reason, we call such a biologically relevant seed alignment a *frame alignment*, by analogy with the skeleton of an evolutionary tree, assuming that the topology of its deepest branches is already well defined since residues in each column are supposed to be consistently and correctly aligned (Fig. 1, stage I). In the second, entirely automated, step, all the remaining homologous sequences are sampled along a decreasing gradient of evolutionary distances and further clustered in order to be added selectively and stepwise, using the closest sequences present in this frame alignment as a template at each step. To continue the tree analogy, building such a *reference alignment* with our entirely automated tool, *Frali* corresponds to the gradual addition of more recent twigs and leafs mainly on the existing deep branches (Fig. 1, stage II). Moreover, securing the first stage allows automation of the process of continuously updating the reference MSA when newly published genomes become accessible, while keeping permanent its accuracy and biological relevance.

2 Methodology

2.1 Stage I: Building the seed (frame alignment)

We regard as representative sequences the few proteins that have been experimentally studied and thus are supposed to be correctly annotated. Optimally, at least one representative sequence of each distant subfamily must be included. Moreover, to assess alignment, we preferentially chose experimentally studied proteins that have been crystallized. Accordingly, the primary amino acid sequences were multiply aligned using Expresso [8]. Whenever the number of sequences with known structures was too low, and/or some subfamilies lacked 3D structure data, we used PSI-Coffee since this is ranked as the most accurate program immediately after 3D structure–based algorithms [3].

Although those automated methods are generally efficient, we always had to review manually the frame alignment obtained so that errors – such as the introduction of indels in structural data –could be avoided. This manual check was made by visualizing the aligned 3D structures using ad hoc tools [9].

2.2 Stage II: From the frame to the reference alignment

Once an optimal seed alignment has been obtained, the remaining homologous sequences can be automatically added to the produced frame to build a reference alignment (Fig. 1, Stage II) using Frali. To maintain a high level of accuracy during the whole process, the addition is made stepwise, as summarized in Fig. 2: clustering the homologues, matching them with their closest partners in the frame alignment, and aligning their hidden Markov model (HMM) based profiles [10].

2.2.1 Preparing a high-confidence reference alignment. To facilitate their targeted addition to the reference alignment, we first clustered the homologues sharing >70% sequence identity over >70% of the length of the shorter matching sequence using the fast and alignment-free CD-hit program [11]. In parallel, their closest homologues prealigned in the frame were likewise clustered. Each cluster was processed through the steps given in Fig. 2:

2.2.1.1 Detecting matching clusters. Since the sequences belonging to such clusters are very close by construction, one of them should reasonably be sufficient to search for matching sequences in the corresponding set of prealigned sequences in the frame. Indeed, although the number of new sequences would seem to be huge, a large fraction of them are actually the n^{th} near identical copy of the same sequence, since they are encoded by different strains of the same species or closely related species. Consequently, instead of comparing each sequence of each cluster to every reference alignment

sequence, we defined one sequence representing each kind of clusters. Thus, the number of representatives increases far more slowly than the raw number of sequences. Representative sequences of each cluster were selected as the longest sequence to avoid accidentally using fragments.

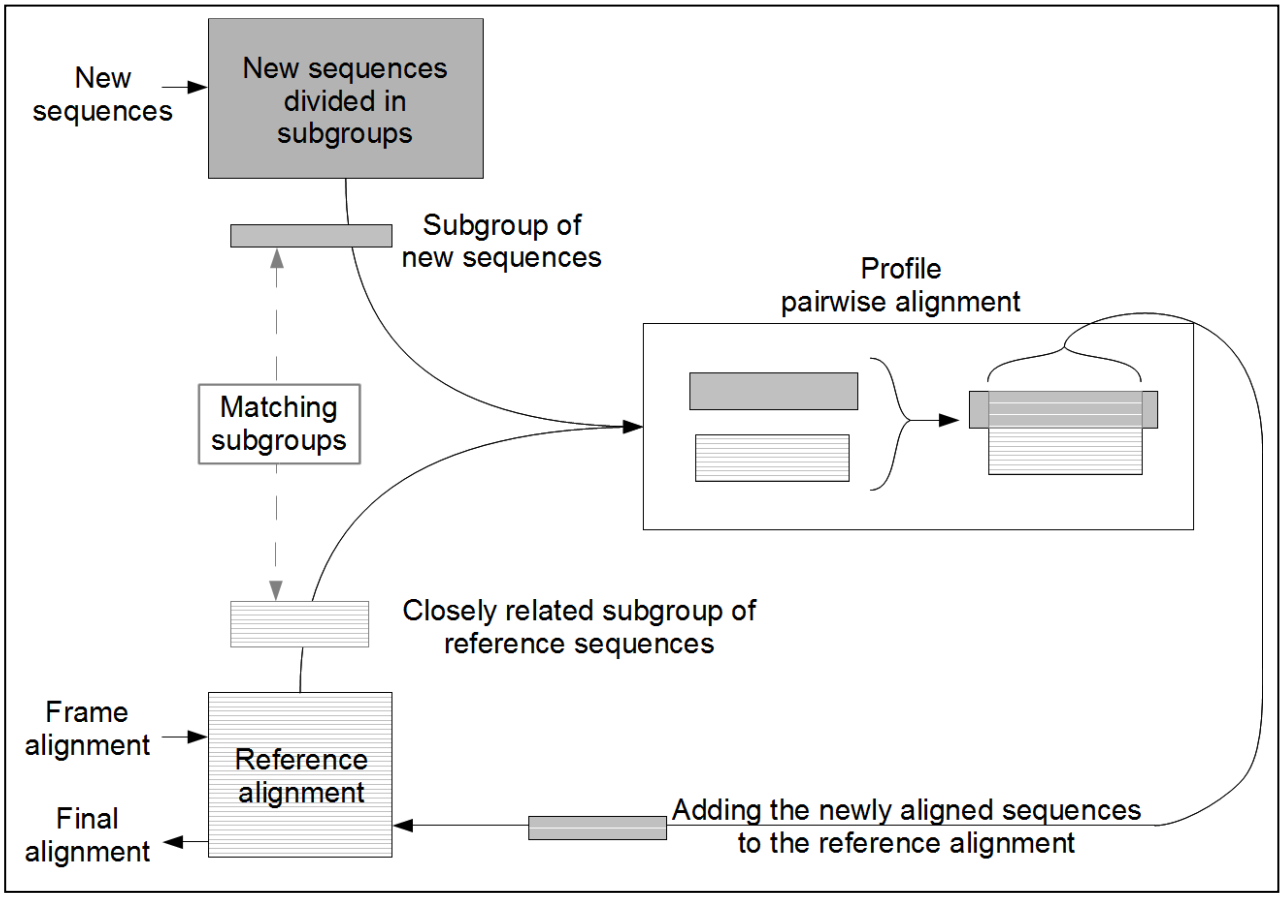


Fig. 2. Outlining the main steps of Frali: new homologous sequences are clustered and each cluster is matched with its closely related subgroup present in the frame alignment by aligning their HMM profiles, allowing their facilitated addition to the reference alignment.

2.2.1.2 Aligning matching clusters. First, the matching frame alignment was stripped of its empty columns before building the HMM profile, and adding them back into the final reference alignment. The new homologues are aligned using Muscle [5], one of the very few methods able to handle large datasets in reasonable times [2]. Although the intrinsic performance of heuristic methods like Muscle is not optimal (discussed in [3]), we ascertained that the elevated level of identity of these highly-related sequences ensures the biological relevance of the obtained MSA. Once matching clusters have been detected using their representative sequences, all their respective sequences were aligned to the prealigned closest frame sequences. Such a progressive addition by subgroups is a crucial trait in our approach. Indeed, it precludes the probable blurring of features specific to each subgroup, such as conserved residues or specific indels (not shown) that would appear in the case of a unique addition of many highly divergent proteins. Noticeably, aligning such groups of closely related homologues allows the further generation of accurate HMM profiles for both the cluster of new sequences under study (HMM_cluster) and the associated cluster of the frame (HMM_frame). The two HMM profiles are then fused using the HHalign program [10]. After addition of each cluster, 2 important points are examined by Frali. (i) Frali extracts selectively the part of HMM_cluster aligning with the HMM_frame by excluding any sequence element located before or after the aligned fraction so as to maximize the efficiency of the HHalign step. This is crucial in discarding the unalignable part (columns absent in the frame) and automatically outlining the homologous segment present in fused proteins. (ii) Frali prevents the misalignment of sequences that are too divergent from the template sequences. Noisy profiles are precluded by impeding the

addition of too distant sequences that will introduce holes of >30 residues. Note that such a safety device does not restrain the addition of sparse natural indels in newly added sequences, since these gaps could be precious phylogenetic markers [27]. Thus, this clusterization step is clearly maintaining the biological relevance when progressively enlarging the frame alignment to the reference alignment, while automated tools would locally damage this relevance (not shown).

2.2.1.3 Improving the HMM alignment and reiterating the whole process for the other clusters. The profile-profile alignment is improved by keeping the accepted indels in the new sequences while reinjecting the common indels that were present in the frame prealignment. This improved cluster alignment is added to the reference alignment. The three steps of the process described above are repeated iteratively for all the other clusters of the set of highly-related sequences, delivering finally a safe reference alignment

2.2.2 Stepwise addition of increasingly distant homologues to the reference MSA. The whole process in Fig. 2 is repeated iteratively while decreasing stepwise the threshold values of sequence identity that are imposed when building the clusters of related sequences to be added, and when matching these clusters to their homologues in the previous reference MSA. These 2 clustering steps are executed once at the beginning of the program and are required for only a few seconds due to the speed of the CD-hit program [11]. Frali progressively processed the homologues found at the 60, 50, 40, and 30% sequence identity cutoffs. Such a stepwise computation of successive new profile-profile alignments is essential in getting a final correct reference MSA, especially when the level of identity becomes too low, while resolving specific problematic cases listed below: (i) Two filters are applied to prevent the introduction of fragments in the reference alignment. First, a maximum length value (which may be defined for each subfamily studied) is imposed as a cutoff before sequence addition to the multiple alignment. A second filter is used after the sequences were aligned, to ensure that the aligned part is complete. This is important, for instance, in the case of multi-domain proteins that are a particular challenge for multiple alignment methods [2]. Since the alignment is done by aligning a query against a template, only the alignable parts of fused proteins will be automatically kept by Frali in the final alignment. The unalignable fragments are set aside in a distinct file that can be read later. (ii) Moreover, we have added a script that detects fused pro-teins where the combined domains of the same protein are homologous to one another (see below, for instance, the case of TrpF and TrpC enzymes). After their detection based on the knowledge of the prealigned frame, these homologous domains are cut and properly aligned during the making of the reference alignment. (iii) Whenever the number of unwanted gaps increases, it might be better to refrain from adding uninformative holes. Keeping the reference alignment as such may prove more stable, since its length would not vary every time an odd sequence appears. Where a significant number of sequences require a common and large gap, the user might consider adding it manually before adding new sequences

3 Implementation of Frali

Frali (http://embg.igmors.u-psud.fr/frali/) is a standalone Perl script package working in a Linux environment with a command-line mode. Frali includes its own modules, and the binary executables needed, such as CD-hit, Blastall, Formatdb, HHalign, and Muscle, provide for both 32 and 64 bit operating systems.

Frali requires 2 main sets of previously computed data: (1) the frame alignment that has been built semi-manually on the basis of expert knowledge (see above), (2) all available homologues that have been collected, as described above. Both inputs are prepared as text files containing FASTA formatted sequences. The output files in FASTA format contain the final updated reference alignment, the leftover sequences that could neither be aligned nor added, and fragments (sequences too small to be added).

Frali can also be used to add directly into the reference alignment new homologous sequences as soon as they are released in public databases. Our choice of defining a representative sequence for each new cluster (see above) allows an acceleration of the process without loss of accuracy. Such a fast and easy update is very helpful for users interested in curation of functional annotation and/or keeping constantly up-to-date phylogenetic trees.

4 Assessing Frali

4.1 Evaluating the accuracy of Frali

We compared the outputs of our 2-step approach

with those of different automated programs, namely ClustalW [13], Dialign [14], Dialign-TX [15], Mafft [16], Muscle [5], Probcons [17], Tcoffee, and 3Dcoffee [18]. Among the benchmark reference alignments described in BAliBASE 3.0 [19] we have utilized the whole package Rev30 made up of 30 aligned families (containing from 24 to 142 sequences), using either the whole sequences or only their homologous regions. Since these RV30 families contain subfamilies with >40% similarity but <20% similarity across the subfamilies, we first applied the psi-CD-hit program [11] to build 10 different seeds for each family by drawing lots among the clusters of its members that share >30% sequence identity. These 600 sets contain 2-15 members (from 2-38% of the total number of family members). Each set was submitted to 2 parallel actions: (i) the sequences extracted from the original BAliBASE alignment were used as a reference seed to which the rest of the homologues were added using Frali; (ii) the full set of all these homologues were submitted to each automated program as unaligned sequences. Since Frali discards the unalignable part of the sequences (Fig. 2 b2), this part (that varies from one seed to the other) was systematically removed before carrying out the reference MSA generated by the automatic tools. This removal was essential to preclude any bias when assessing the obtained reference alignments by measuring the number of correctly aligned residue pairs divided by the number of aligned residue pairs in the true alignment (score SP) and the number of correctly aligned columns divided by the number of columns in the true alignment (score TC), as defined in Thompson et al. (2005).

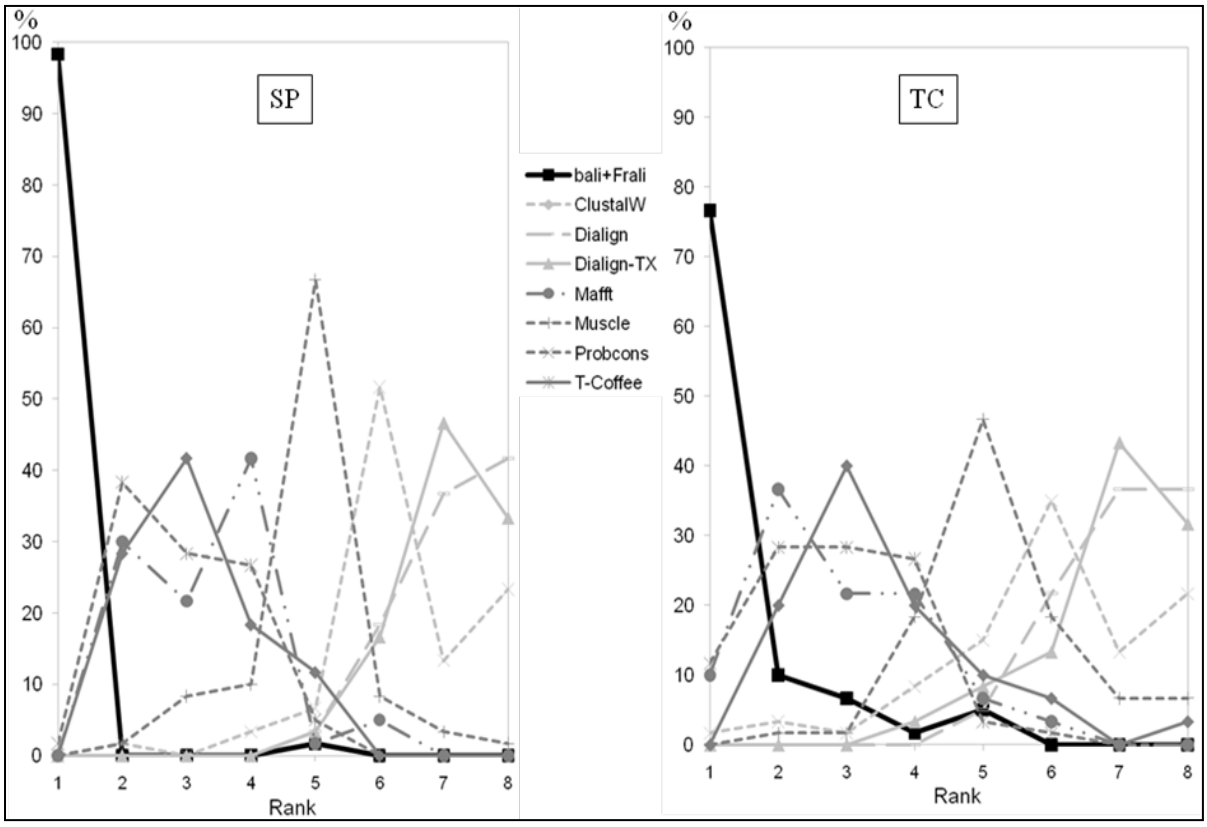


Fig. 3. Ranking the accuracy of Frali using protein alignment benchmarks. For each BAliBASE family 10 different clusters of sequences were built that have been multiply aligned using either our 2-step approach (bali+Frali) or various automated programs listed between panels SP and TC. For comparison, exactly the same portions of each sequence included in each set have been used to build the final MSA. We computed for each set the rank of each program using 2 BAliBASE scores, namely SP (left panel) and TC (right panel), and we computed for each family the median of the ranks in its 10 respective sets. Left and right panels show the percentage of families where each program has been ranked in position 1 to 8.

To gauge each method, we first ranked the SP and TC scores of each program for each set of each family and we further classified each program by computing the median of these 10 ranks for both scores in each family (Fig. 3). Our approach appears to be significantly more accurate than the tested

automated programs (Fig. 3) since it is ranking in the first position in 76.67% of the analyzed families regarding the TC score and in 98% of the tested families when measuring the SP score. Note that when automated programs perform better than Frali, namely 12.67% with Probcons, 10% with Mafft, and 1.67% with ClustalW in the case of the TC score, these BAliBASE families were made of closely related sequences. Moreover, in those cases, Frali is generally ranked second, giving a very slightly lower score.

4.2 Testing the biological relevance of Frali using a challenging family

Two models describing the possible evolution of enzyme activities were experimentally validated a decade ago when a gene encoding the TrpF activity was obtained by transforming either the gene encoding the HisA activity [19] according to the patchwork model [20] or the gene encoding the previous TrpC step [21] according to the retrograde model [22]. Moreover, another retrograde case was previously described since HisA was found to be homologous to its next step HisF [23,24]. Besides, TrpA appears to be distantly related to TrpC and TrpF (unpublished data). Thus, five genes encoding TIM-barrel proteins – TrpA, TrpC, TrpF, HisA, and HisF - are found to form a family of homologues that are probably very ancient. Indeed, the sequence identity separating these exhaustively studied proteins was found to be low (25% separating HisA from HisF) to extremely low (only 11% between HisA and TrpF and 13% between HisF and TrpF according to [25]), but their X-ray structures are superimposable. Thus, these remotely related structural homologues appear to be a challenging test case for analyzing the relevance of Frali.

Since the 3D structures of the majority of these enzymes have been determined, we could build a frame MSA with 19 sequences using either Expresso [8] or Muscle [5]. Unsurprisingly, these two automated programs gave unsatisfactory alignments and the deduced trees built using the FastTree2 program display poor biological relevance (not shown). As described above, we improved this seed alignment to a faithful alignment after manual expert edition using Swiss-PdbViewer 4.0 [9]. The tree reconstructed automatically using Muscle [5] and Expresso [8] were biologically less relevant than the ones obtained from the manually built frame MSA (Fig. 3, left panel) since their HisA and HisF subtrees were not monophyletic and branch with TrpF sequences (not shown). Moreover, their relative branch length and topology were longer than that of the tree built from the frame alignment taken as a reference. Indeed, the K tree scores [26] of Muscle and Expresso trees are 1.25787 and 0.63350, respectively. Fig. 3 further shows how Frali allows building progressively a reference alignment with selected addition to the frame (left panel) of the homologues displaying first at least 40% identity (central panel) and then the rest of the 3229 more distant homologues (right panel). The deduced phylogenetic tree keeps the same skeletal structure already observed in the frame alignment, each subfamily becoming just more and more burgeoning.

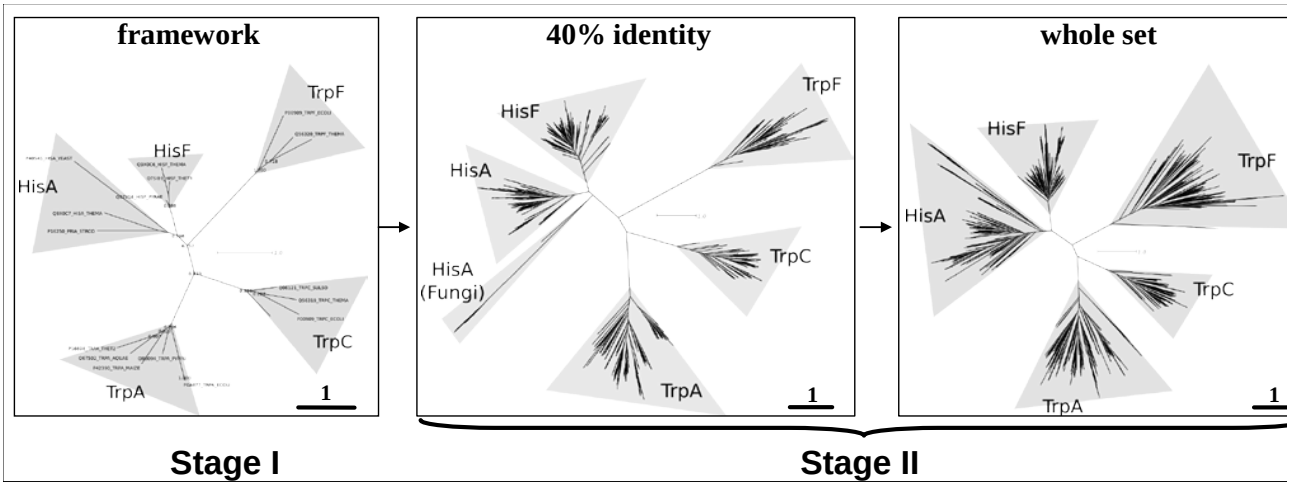


Fig. 4. Progressive addition of newly related sequences to the frame tree (left panel) reconstructed from a manually improved MSA . Trees were built using the FastTree2 program [12]. Central panel shows how the newly added sequences are added nicely as twigs specific to each subfamily on the conserved frame of the whole tree. Right panel confirms this selective addition on an invariant ancient tree skeleton with a concomitant shortening of the deepest branches.

5 Discussion

Frali has been designed to help the biologist escape various methodological and conceptual difficulties when building multiple alignments of large and diverse arrays of homologues that can be very distant. The proposed approach has a cost, since it requires a preliminary manual editing of the MSA of a limited number of experimentally well-characterized proteins that stand for the various subfamilies of such arrays. This limited number of seed sequences could be as low as 5% of the total number of family members. Once such a solid basis is established, the whole alignment can be obtained very rapidly by using the completely automated Frali program. This reasonable effort of manual editing is rewarding in the end since it can guarantee getting a reference MSA that is both accurate and biologically relevant. This is mainly due to our strategy of progressive addition of new homologous proteins that have been sampled by tight clustering, defining a high similarity to a few of the prealigned sequences in the frame alignment. This careful handling of the sequences during the profile-profile step and the strict treatment of the indels helps maintain the accuracy of the obtained reference alignment, as shown in comparative studies with automated programs on the same benchmarking data (Fig. 3). Noticeably, contrarily to the case of completely automated one-step methods, the biologist will keep mastering the intricacies of the process of multiply align complex families of homologous sequences at each step of the Frali approach, even when they are highly dissimilar.

Our tool presents several decisive advantages over other methods. (i) Whatever the present and future level of flooding of newly released genomic sequences, we guarantee the accuracy of the MSA since we start with a high level of truthfulness at the step of the frame alignment, and we keep it unabated when adding stepwise and gradually the whole set of the other homologues. (ii) Our procedure is fast, its rate being linearly proportional to the increase in the total number of sequences to be aligned. (iii) Frali resolves instantaneously difficult cases such as multi-domain and/or fused proteins without any prior detection or treatment. (iv) The opening of too large holes is prevented by our gradual and stepwise procedure, but the possibility of introducing a limited number of gaps is kept since they could be valuable phylogenetic markers [27]. (v) Phylogenetic trees derived from MSA generated with Frali systematically display a better topology and a shorter length than those derived using one-step automated tools.

In addition, the full reference MSA may be updated at any time while keeping its accuracy and biological relevance. Indeed, addition of newly published homologues takes a few seconds and is highly precise. Therefore, Frali allows effortlessly the last update of a phylogenetic tree of a large and complex family to be generated at anytime. Note, however, that the occurrence of representatives of a completely new sub-family could require a supplementary step before their addition to the reference alignment.

Acknowledgements

This work was funded by the CNRS (UMR 8621), the PPF ‘Bioinformatique et BioMathématique’ of the Université Paris-Sud and the Agence Nationale de la Recherche (ANR-05-MMSA-0009 MDMS_NV_10). M.B. is a PhD student supported by the French Ministry of Research.

References

[1] Pei J. Multiple protein sequence alignment. Curr Opin Struct Biol. 18:382-386, 2008

[2] Pirovano W, Heringa J. Multiple sequence alignment. Methods Mol Biol. 452:143-161, 2008

[3] Kemena C. Notredame C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics 25:2455-2465, 2009

[4] Edgar:R.C. and Batzoglou, S. Multiple sequence alignment. Curr. Opin. Struct. Biol.16:368–373, 2006

[5] Edgar R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113, 2004

[6] Patterson C Homology in classical and molecular biology:Mol. Biol. Evol. 5:603-625, 1988

[7] Descorps-Declère S Lemoine F. Sculo Q Lespinet O and Labedan B. The multiple facets of homology and their use in comparative genomics to study the evolution of genes:genomes:and species. Biochimie 90:595-608, 2008

[8] Armougom F:Moretti S:Poirot O:Audic S Dumas P Schaeli B Keduas V Notredame C. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res. 34(Web Server issue):W604-608, 2006

[9] Guex N. Peitsch M.C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis 18:2714-2723,

1997

[10] Eddy SR. Profile hidden Markov models. Bioinformatics 14:755-763, 1998

[11] Li W Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659, 2006

[12] Price M.N. Dehal P.S. and Arkin A.P. FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. Mol Biol Evol 26:1641-1650, 2009

[13] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 2:4673-4680, 1994

[14] Morgenstern,B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics, 15:211–218, 1999

[15] Subramanian AR, Kaufmann M, Morgenstern B DIALIGN-TX: greedy and progressive approaches for the segment-based multiple sequence alignment. Algorithms Mol. Biol., 3:6, 2008

[16] Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. PROBCONS: Probabilistic Consistency-based Multiple Sequence Alignment. Genome Research, 15:330-340, 2005

[17] O'Sullivan,O. Suhre K, Abergel C, Higgins DG, Notredame C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J. Mol. Biol., 340:385–395, 2004

[18] Thompson,J.D., Koehl,P., Ripp,R. and Poch,O. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins, 61:127–136, 2005

[19] Jürgens C Strom A Wegener D Hettwer S Wilmanns M Sterner R Directed evolution of a (beta alpha)8-barrel enzyme to catalyze related reactions in two different metabolic pathways. Proc Natl Acad Sci USA. 97:9925-9930, 2000

[20] Jensen RA Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409–425, 1976

[21] Altamirano M. M. Blackburn J.M. Aguayo C. Fersht A.R. Directed evolution of new catalytic activity using the alpha/beta-barrel scaffold Nature 403:617–622, 2000

[22] Horowitz:N.H. On the evolution of biochemical syntheses. Proc. Natl. Acad. Sci. USA 31:153–157, 1945

[23] Fani R Mori E Tamburini E Lazcano A. Evolution of the structure and chromosomal distribution of histidine biosynthetic genes. Orig Life Evol Biosph 28:555-570, 1998

[24] Lang D Thoma R Henn-Sax M Sterner R Wilmanns M Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. Science 289:1546–1550, 2000

[25] Leopoldseder S Claren J Jürgens C Sterner R Interconverting the catalytic activities of (ba)8-barrel enzymes from different metabolic pathways: sequence requirements and molecular analysis. J Mol Biol 337:871–879, 2004

[26] Soria-Carrasco V Talavera G Igea J Castresana J. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics. 23:2954-2956, 2007

[27] Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 320:1632-1635, 2008

Functional and structural disorder: comparative genomics and genetic interactions distinguish functional roles of disorder

Jeremy BELLAY[2*], Sangjo HAN[1*], Magali MICHAUT[1*], Gary D. BADER[1], Chad L. MYERS[2] and Philip M. KIM[1]

[1] Terrence Donnelly CCBR, University of Toronto, 160 College Street, M5S 3E1, Toronto, Ontario, Canada
{sangjo.han, magali.michaut, gary.bader, pm.kim}@utoronto.ca
[2] Department of Computer Science and Engineering, University of Minnesota, MN 55455, Minneapolis, USA
{bellay, cmyers}@cs.umn.edu
[*]These authors contributed equally to this work.

Abstract *Intrinsically disordered regions are common in many proteins and have been associated with many diseases and functions. In this study, we distinguish different types of intrinsic disorder using genetic interactions and comparative genomics.*

Keywords Disorder, Intrinsically Disordered Protein, Conservation, Genetic Interaction.

1 Introduction

Intrinsically disordered regions are common in many proteins, especially in higher eukaryotes [1]. Intrinsically disordered proteins (IDPs), which have a large fraction of disordered residues, have been associated with many diseases and functions, but there is still much active investigation about their cellular roles [2]. In this study, we distinguish different types of intrinsic disorder using genetic interactions [3] and comparative genomics.

2 Results

2.1 Genetic Interaction Network Identifies Functional Disorder

We first observe that genes that have numerous genetic interactions (hubs) often tend to encode proteins that have a higher percentage of disordered residues (Fig 1).

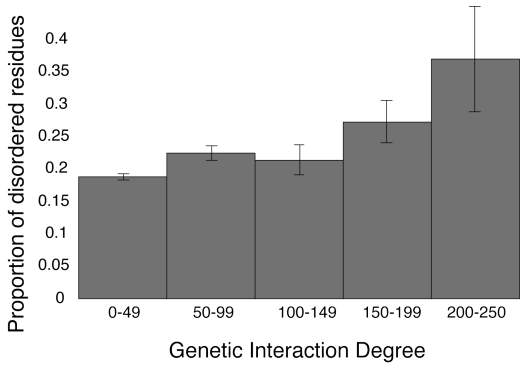


Fig. 1. The degree of a gene in the genetic interaction network is correlated with the % of disorder of the gene product. Genetic hubs tend to encode disordered proteins.

Interestingly, IDPs are split into two groups: among the hubs, the level of disorder tends to be highly associated with multifunctionality, whereas the IDP non-hubs do not exhibit this correlation (Fig 2). The IDP genetic network hubs appear to be responsible for previous associations of disorder with signaling and also show strong enrichment for so-called date hubs [4] and single-interface hubs [5].

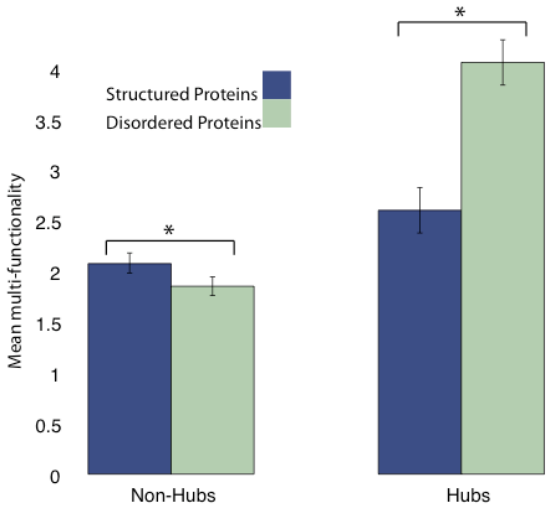


Fig. 2. Among the genetic hubs, disorder is associated with multifunctionality whereas is it not the case for disordered proteins that are encoded by non hub genes.

2.2 Defining Conserved Disorder

We hypothesized that we could further distinguish different forms of disorder with different functional contexts by examining evolutionary properties. Specifically, we investigated which disordered regions were also disordered in orthologous proteins

across the yeast clade. Intriguingly, we found strong evidence for "conserved disordered" regions across the yeast clade, indicating selection on disorder as a structural property often with little or no constraints on the actual sequence.

2.3 Phosphorylation Sites are found in Regions of Conserved Disorder

We found these regions of conserved disorder are strongly associated with proteins harboring many linear motifs and are specifically predictive of phosphorylation sites (Fig 3), indicating their critical role in signaling. In contrast, disordered regions that are not conserved do not exhibit strong correlation with these features and likely correspond to disorder that is not functional.

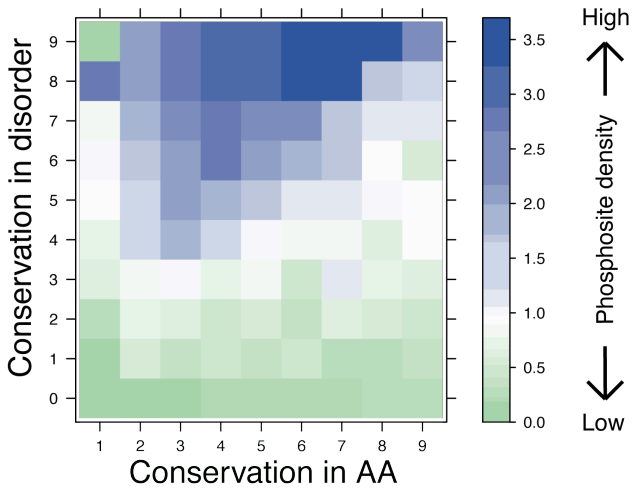


Fig. 3. The density of phosphorylation sites is represented for regions of various degrees of conservation at the amino acids (AA) and disorder levels. The blue region shows that the conservation of disorder in a region is associated with a high density of phosphorylation sites whereas there is no such association with the conservation of specific amino acid sequence.

2.4 Amino Acid Conservation distinguishes Functional Roles of Conserved Disorder

Finally, we found that regions of conserved disorder with quickly evolving sequences are functionally distinct from conserved disorder where the underlying amino acid sequence is highly constrained. This class of disordered proteins has markedly different signatures in a variety of physiological and functional data and appears to be associated in RNA/protein binding and protein folding (Fig 4).

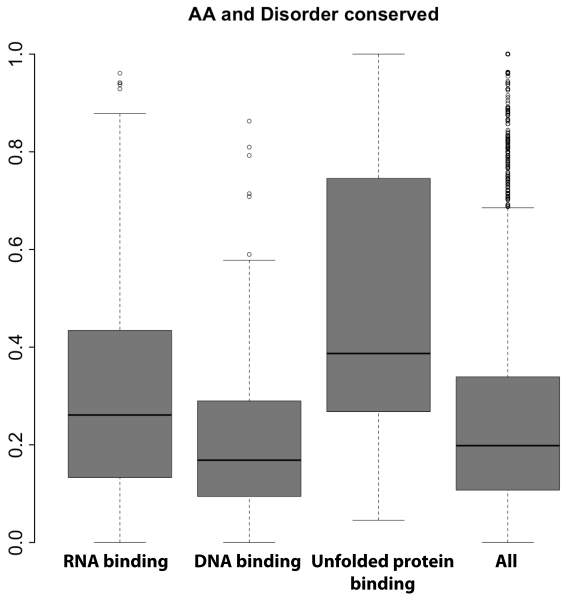


Fig. 4. The percentage of positions with amino acid and disorder conservation shows significant differences in various families of proteins.

3 Conclusions

In summary, we split the protein disorder along two axes i) conservation of disorder, which separates the functional from non-functional disorder. This corresponds to the distinction between hubs and non-hubs in the genetic interaction network ii) conservation of amino acids in disordered regions, which separates signalling from RNA binding and protein folding/chaperone activity.

References

[1] J Gsponer, ME Futschik, SA Teichmann, MM Babu, Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science, 322(5906):1365-1368, 2008

[2] AK Dunker, CJ Oldfield, J Meng, P Romero, JY Yang, JW Chen, V Vacic, Z Obradovic, VN Uversky, The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics, 9 Suppl 2S1, 2008

[3] M Costanzo, A Baryshnikova, J Bellay, Y Kim, ED Spear, CS Sevier, H Ding, JLY Koh, K Toufighi, S Mostafavi *et al*, The genetic landscape of a cell. Science, 327(5964):425-431, 2010

[4] J-DJ Han, N Bertin, T Hao, DS Goldberg, GF Berriz, LV Zhang, D Dupuy, AJM Walhout, ME Cusick, FP Roth *et al*, Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 430(6995):88-93, 2004

[5] PM Kim, A Sboner, Y Xia, M Gerstein, The role of disorder in interaction networks: a structural analysis. Mol Syst Biol, 4179, 2008

Computational analysis of the dynamics of logical regulatory graphs

Duncan Berenguier[1,2], Claudine Chaouiya[1,3], Élisabeth Rémy[2] and Denis Thieffry[1,4,5]

[1] TAGC, U928 INSERM, 163, Avenue de Luminy, 13009, Marseille, Cedex 09, France
berenguier@tagc.univ-mrs.fr
[2] IML, UMR6206 CNRS, Avenue de Luminy, 13009, Marseille, Cedex 09, France
remy@iml.univ-mrs.fr[2]
[3] Instituto Gulbenkian de Ciência, Oeiras, Portugal
chaouiya@igc.gulbenkian.pt[2]
[4] IBENS - CNRS UMR 8197 / INSERM U1024, Ecole Normale Supérieure, Paris, France
thieffry@ens.fr
[5] CONTRAINTES, INRIA Paris-Rocquencourt, Le Chesney, France

Keywords Regulatory networks, logical models, state transition graphs, attractors.

Summary

The dynamical analysis of large biological regulatory networks requires the development of scalable mathematical modelling methods. Following the approach initially introduced by R. Thomas, we formalise the interactions between the elements of a network in terms of discrete variables, functions and parameters [1]. This approach has been implemented into the software *GINsim*, which enables the definition of *logical regulatory graphs* representing regulatory components (nodes), interactions (arcs) and rules (parameters). Model simulations with *GINsim* result in *state transition graphs,* which represent the temporal behaviour of wild type or mutant regulatory networks [2]. We are particularly interested in asymptotic behaviours, which correspond to terminal strongly connected components or *attractors* in the state transition graph.

For complex networks, the explicit construction of state transition graphs can be cumbersome or even intractable. Therefore we developed computational strategies to cope with this problem, including an algorithm enabling the determination of stable states directly from the model (without computing the state transition graph) [3], and a reduction approach preserving essential dynamical properties [4].

Here, we propose an algorithm to compact state transition graphs on the fly. The result of a simulation is compressed into a hierarchical graph. With this intent, we consider the strongly connected components (SCC) of the state transition graph which are either trivial attractor (terminal SCC consisting in a single node), complex attractor (terminal SCC of cardinal greater than 2), or complex transient components (non terminal SCC of cardinal greater than 2). All the other nodes are trivial transient states (non terminal SCC of cardinal equal to 1). We define σ an application returning, for each trivial transient state, the set of SCCs (complex transient, complex or trivial attractor) reachable from it. The trivial transient states with the same image by σ and connected by a *transient path* are then grouped into a single basin of attraction. The SCCs and the basins of attraction define the nodes of our hierarchical representation of state transition graphs.

Our methodology is based on a modified version of Tarjan's algorithm [5]. During a depth first search, this algorithm associates an index to each node newly encountered, and uses hierarchical properties of these indexes to find the cycles. Since we compress the components in the course of computing, we cannot keep track of these indexes after compaction. When a new strongly connected component (or a basin) is found, each of its states is compacted along with a reference to this strongly connected component in a common decision diagram. Ultimately, this diagram holds all the states and the components they belong to.

An alternative approach would be to derive the structure of the state transition graph directly from the logical regulatory graph. This approach has been successfully applied to characterise stable states [3]. We are now working on an extension of this approach to other kinds of attractors, using recent mathematical results connecting the presence of positive or negative *regulatory circuits* in the regulatory graph with the occurrence of multiple attractors or dynamical cycles in the state transition graph [6].

Acknowledgements

Duncan Berenguier is supported by a PhD grant from the French Ministry of Research and Technology. Our research is further supported by the ANR SYSCOMM CALAMAR project (ANR-08-

SYSC-003), the EU FP7 APO-SYS project, and the the Belgian Science Policy Office (IAP BioMaGNet).

References

[1] R. Thomas, D. Thieffry and M. Kaufman. Dynamical behaviour of biological regulatory networks. I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.*, 57: 247-276, 1995.

[2] A. Naldi, D. Berenguier, A. Fauré, F. Lopez, D. Thieffry, and C. Chaouiya. Logical modelling of regulatory networks with GINsim 2.3. *Biosystems*, 97:134-139, 2009.

[3] A. Naldi, D. Thieffry and C. Chaouiya. Decision diagrams for the representation and analysis of logical models of genetic networks. *Lect. Notes Comput. Sci.*, 4695:233-247, 2007.

[4] A. Naldi, E. Remy, D. Thieffry and C. Chaouiya. A reduction of logical regulatory graphs preserving essential dynamical properties. *Lect. Notes Bioinfo.* 5688:266-280, 2009.

[5] R. Tarjan. Depth-first search and linear graph algorithms. *SIAM J. Comput.* 1:146-160, 1972.

[6] E. Remy, P. Ruet and D. Thieffry. Positive or negative regulatory circuit inference from multilevel dynamics. *Lect. Notes Control. Inf. Sci.* 341:263-270, 2006.

A Rendering Method for Small Molecules up to Macromolecular Systems: HyperBalls Accelerated by Graphics Processors

Matthieu CHAVENT[1], Antoine VANEL[2], Bruno LEVY[3], Bruno RAFFIN[4], Alex TEK[5] and Marc BAADEN[5]

[1]CEA, DAM, DIF, 91297 Arpajon, France.
chaventm@ocre.cea.fr
[2]ID-IMAG, CNRS/INPG/INRIA/UJF, Grenoble, France.
antoine.vanel@inrialpes.fr
[3]Equipe ALICE, Nancy Université, LORIA/INRIA Nancy Grand-Est, 54506 Vandoeuvre-les-Nancy Cedex, France.
Bruno.Levy@inria.fr
[4]ID-IMAG, CNRS/INPG/INRIA/UJF, Grenoble, France.
bruno.raffin@imag.fr
[5]Institut de Biologie Physico-Chimique, Laboratoire de Biochimie Théorique, CNRS UPR 9080, 13, rue Pierre et Marie Curie, F-75005 Paris, France.
{tek,baaden}@ibpc.fr

Abstract *We introduce a new type of representation named* HyperBalls *that links atom spheres using hyperboloids. This graphical rendering method, based on the GPU ray-casting technique, accurately and efficiently represents molecules ranging from a few atoms up to huge macromolecular assemblies with more than 500,000 atoms.*

Keywords GPU computing; Cg ray-casting; improved Ball & Stick, Licorice representations; HyperBalls;

1 Introduction

Different types of representations exist to draw molecules, ranging from the most simple ones such as van der Waals space-filling spheres depicting atom positions to more sophisticated ones such as molecular surface representations that define the overall macromolecular shape. In addition to these two extreme cases, it is possible to define intermediates such as Ball & Stick or Licorice models depicting covalent bonds or ribbon metaphors for representing protein backbone structures. Here, we introduce another depiction, complementary to these well known representations, that could be particularly interesting to represent a continuous dynamic evolution of bonds. This contrasts with existing static representations provided by Ball & Stick or Licorice models that suffer from an "all or nothing" dilemma. We have named this new representation *HyperBalls*, as it is composed of a sphere depicting the atoms linked by hyperboloids. This implicit *HyperBalls* surface relates to MetaBalls, which also provide an analytical expression, but is based on simpler second order equations. The representation using hyperboloid primitives can be adapted to highlight the continuous evolution of bonds – passing from a one sheeted hyperboloid to a two sheeted hyperboloid – and can furthermore encapsulate classical types of representations such as space filling, Ball & Stick or Licorice models. *HyperBalls* hence provide a generic unified molecular representation and is well adapted to render coarse-grained models, even further increasing the number of molecules that can be represented.

Recently, we have developed a method to visualize Molecular Skin Surface using the ray-casting technique on GPUs [1]. Here, we extend this method to implement the *HyperBalls* representation.

2 Results

The *HyperBalls* representation is well suited to depict several different molecular metaphors that were, until now, quite difficult to define precisely. In this part, we will briefly describe several examples.

2.1 Visualizing the dynamic evolution of non-covalent bonds

Visualization of non-covalent bonds in general, and of hydrogen bonds in particular, is a specific point of interest of the *HyperBalls* method. Classically, this particular type of molecular interaction is represented as dashed lines. Another possibility is to visualise interactions in 2D to clarify the scene and focus on a point of interest. A major limitation of these representations is that no clues about the interaction strength are provided: either an interaction exists, or it does not. Using *HyperBalls*, it

is possible to parameterize the representation in order to observe a change of the hyperboloids as a function of the distance between interacting atoms or molecules (see Fig. 1). This metaphor would be particularly useful to depict the temporal evolution of non-covalent bonds to better understand phenomena such as the formation of hydrogen bonds between water molecules.

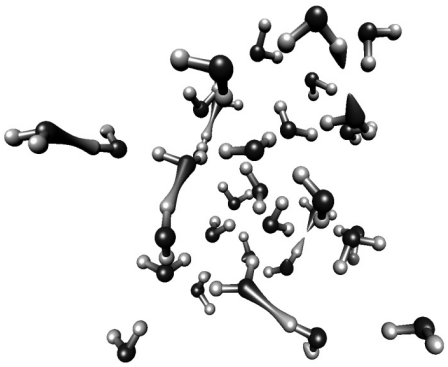


Fig. 1. Hydrogen bond formation emphasized using *HyperBalls*.

2.2 Visualizing ion coordination

Similarly to hydrogen bonds, the coordination between ions and surrounding molecules can be rendered as cylinders or lines. The use of the *HyperBalls* representation provides another perception of such bonds and can be adapted to highlight the ion's coordination environment (see Fig. 2). This can be particularly useful to illustrate asymmetries in the coordination sphere.

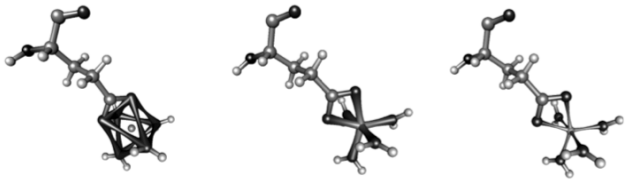


Fig. 2. Ion coordination visualization.

2.3 Rendering huge molecular assemblies

The *HyperBalls* representation is well suited to depict reduced models of big molecular systems such as coarse-grained models or elastic spring networks. Given that the thickness of the bonds can be varied as a function of the distance between atoms, the *HyperBalls* representation provides some distinct advantages for depicting springs or coarse-grained bond behaviour (data not shown).

Futhermore, the use of GPU ray-casting allows us to visualize huge assemblies at real time frame rates. For example, it is possible to visualize virus capsids containing over 560,000 atoms with a frame rate of 35 fps (see Fig. 3), which cannot be achieved with common molecular viewers such as VMD or Pymol, on the same hardware. It is also possible to visualize molecular dynamics trajectories of macromolecular assemblies containing more than 300,000 atoms with an interactive frame rate of 17 fps.

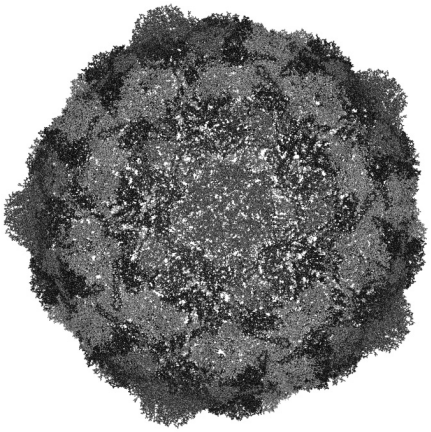


Fig. 3. Virus capsid (~560,000 atoms) visualization.

3 Conclusion

In this paper, we have introduced a new molecular representation called *HyperBalls*. This new molecular representation is well adapted to display non covalent bond changes or coarse-grained models. Furthermore, we used the ray-casting technique on GPUs for accurately and efficiently visualizing the *HyperBalls* representation, even for huge assemblies. A simple viewer program should be available soon, but the final goal is to implement such algorithms in well known molecular viewer programs such as VMD and make them widely available.

Acknowledgements

This project was funded by the French Agency for Research (FVNANO grant ANR-07-CIS7-003-01; http://www.baaden.ibpc.fr/projects/fvnano). M. Chavent thanks J.P. Nominé (CEA, DIF) for insightful comments.

References

[1] M. Chavent, B. Levy, B. Maigret. MetaMol: high-quality visualization of molecular skin surface. J Mol Graph Model 2008, 27(2), 209-216.

Lineage-specific orthologous gene loss and pseudogenisation, automated analysis in Metazoans

Jacques Dainat[1], Julie D. Thompson[2], Olivier Poch[2], Pierre Pontarotti[1], Philippe Gouret[1]

[1] LATP - Evolution Biologique & Modélisation, UMR 6632 CNRS, Université de Provence, Site St Charles - case 19 Place Victor Hugo, 13331, Marseille Cedex 03, France
jacques.dainat@etu.univ-provence.fr

[2] Laboratoire de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, BP 10142, 67404, Illkirch, France
julie.thompson@igbmc.fr;Olivier.Poch@igbmc.u-strasbg.fr

Keywords Automation, gene loss, genome, ortholog, phylogeny, pseudogene.

1 Introduction

Pseudogenes were first discovered in 1977, in the Xenopus laevis genome. Since, different studies have reported pseudogenes in several organisms, ranging from bacteria to human. Two different types of mechanisms are involved in the emergence of pseudogenes: First, pseudogenes can appear through a process of genomic material duplication. This phenomenon mostly arises by a retrotransposition because they are derived from a mature mRNA product and it lacks the upstream promoter region and the introns of normal genes [1]. These pseudogenes called “processed pseudogenes” are often considered “Dead on Arrival”, as they become non-functional immediately upon the retrotransposition event [2]. The second mechanism consists in the drift from existing functional genes to pseudogenes through the accumulation of deleterious mutations. This phenomenon mostly arises after a duplication event, who generates redundant paralogous copies of a gene. One of outcomes in the evolution of duplicated genes is that one copy becomes silent through the accumulation of degenerative mutations [3]. However, it exists universally conserved genes that also undergo such a gradual erasure of the sequence. A well-established gene may have an essential function, and its disappearance may reflect an important phenomenon. Such a gene loss could be adaptive: that gene could either encode a no more useful function, or some other genes in the genome could fulfill that function. The availability of complete genomes sequences allows to investigate gene losses at a larger scale, and to consider co-losses in different lineage. So, to perform automatically these analyses, we developed the module “GeneLoss” that is part of the multi-agent system “Dagobah”. The framework “Dagobah” is a set of autonomous softwares running in parallel, communicating between each other and with external software platforms (Ensembl [4], NCBI [5], Figenix [6]), and sharing persistent results. “Dagobah” aims to perform, completely automatically, the prediction and the phylogenetic placement of all the genetic events that occurred during the evolutionary history of genomes. Within “Dagobah”, the ”GeneLoss” agent is dedicated to specifically study the gene deletion and the pseudogenisation processes. Finally, our “GeneLoss” module is also be able to find intact but un-annotated genes.

2 Strategy of “GeneLoss” module

The aim of “GeneLoss” is to re-annotate genes in a specific lineage from a reference gene. From a robust multiple sequences alignment built from a set of species, we build a phylogeny through a pipeline of the automated platform Figenix. Phylopattern [7] identifies an orthologous group, which contains the query protein. The comparison of orthologous species with the list of species under investigation, allows identifying species where the orthologous are missing. We verify the ortholog absence using the “TBlastN” algorithm on the Ensembl database of the species concerned. When a hit is found, we check by new phylogenetic analysis if the hit is ortholog to the group defined before. If no ortholog is found, we conclude to a gene loss. If an ortholog is found, we investigate whether it is a mis-annotated gene, or a pseudogene. Analyses are first done at the protein level. To do this we extract a piece of DNA containing the ortholog signal found and we try to predict the most similar protein with GenePredix [6]. Then, according to an identity threshold, we test the length and the similarity of the homolog sequences (hit and prediction). When the ortholog protein sequence found is under the threshold, the study remains at protein level. We check if the protein conservation is consistent with time of divergence with the last ancestor known which encodes the protein. The consistence is calculated by a mean of identity divergence by million of years according to

the known orthologous protein of species studied. If the signal is consistent, we conclude to the putative existence of that orthologous gene, otherwise we conclude to its pseudogenisation. When the ortholog sequence found is not too divergent, we perform a study at the genomic level. We align with "LaganM" [8] the DNA sequence with ancestral sequence reconstructed by "Ortheus" [9]. Then, we search all possible mutation events, comparing the target sequence to the ancestral sequence. If no degenerative mutation exists, we conclude the putative existence of the gene. In the other case we conclude to the pseudogenisation of the gene. The last step is to summarize the results. Each species has a character for the gene of interest: present, apogene (for putatively present), pseudogene, or lost. The Sankoff parsimony [10] is used to highlight the ancestral and derived characters. At the end of study, these information and the genomic events are pinpointed on a tree of life.

3 Results and Discussion

To show the efficiency of "GeneLoss" module, we have benchmarked it with more than twenty cases of gene loss described in literature. Among them there are gene losses in the primate lineage [11], in the hominidae lineage [12,13], in the human lineage [14] and in the dog lineage [15]. For each of this studies, we found the same results than those published. Furthermore, for most of them, we identified new interesting information and further details, which allowed us to refine the previous descriptions. Here we will take to example the gene coding to acyltransferase 3 (*Acyl3)* protein [12], described by the authors as a pseudogene in Homo and Pan, due to a nonsense mutation in exon seven appeared after the divergence of gorillas from the human lineage and before the human-chimp split. With "GeneLoss" we found the mutation already described, and many others non-described mutations. A splice mutation in the ancestor of Hominidae seems to be the first event leading to the process of pseudogenisation. Four nonsense mutations and one insertion of four bases occur after this event in different Hominidae lineage. In addition, our analysis also revealed a loss of the *Acyl3* gene in the Gallus gallus genome, which has never been described yet.

4 Conclusion

"GeneLoss" is the first tool able to study gene loss and the pseudogenisation automatically. We demonstrated its accuracy and its efficiency. Currently, a large scale analysis is undertaken in the human lineage case.

Acknowledgements

Dainat J. is supported by a PhD Student fellowship from the French Ministry of Research. This work is supported by the French National Research Agency [EvolHHupro: ANR-07-BLAN-0054].

[1] Vanin EF: Processed pseudogenes: characteristics and evolution. Annu Rev Genet, 19: 253-272, 1985.
[2] Graur D, Shuali Y, Li WH: Deletions in processed pseudogenes accumulate faster in rodent than in humans. J Mol Evol., 28: 279-285, 1989.
[3] Lynch M, Coney JS: The evolutionary fate and consequences of duplicate genes. Science, 290: 1151-1155, 2000.
[4] www.ensembl.org
[5] www.ncbi.nlm.nih.gov
[6] Gouret P, et al: FIGENIX: Intelligent automation of genomic annotation in a new software platform. BMC Bioinformatics, 6:198, 2005.
[7] Gouret P, et al: PhyloPattern: regular expressions to identify complex patterns in phylogenetic trees. BMC Bioinformatics, 10:298, 2009.
[8] Brudno M, et al: NISC Comparative Sequencing Program, Green ED, Sidow A, Batzoglou S, LAGAN and Multi-LAGAN: Efficient tools for large scake multiple alignment of genomic DNA. Genome Research, 13: 721-731, 2003.
[9] Paten P, et al: Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Research, 18: 1829 1843, 2008.
[10] Sankoff D: Minimal Mutation Trees In Sequences. SIAM J Appl Math, 28: 35-42, 1975.
[11] Yuriko Ohta and Morimitsu Nishikimi: Random nucleotid substitution in primate nonfunctionnal gene for L-gulono-gamma-lactone oxidase, the missing enzyme in L-ascorbic acid biosynthesis. Biochimica et Biophysica Acta, vol. 1472: 408-411, 1999.
[12] Zhu J, et al., Comparative Genomics Search for losses of long-established genes on the human Human lineage. PloS Compute Biol, 3:e247, 2007.
[13] Derouet D, et al: Neuropoietin, a new IL-6-related cytokine signaling through the ciliary neurotrophic factor receptor. Proc Natl Acad Sci USA, 101: 4827-4832, 2004.
[14] Derrien T. et al: Revisiting the missing protein coding gene catalog of the domestic dog. BMC Genomics, 10:62, 2009.
[15] Nishikimi M, et al: Cloning and chromosomal mapping of the human nonfunctionnal gene for L-Gulono-γ-lactone Oxidase, the enzyme for L-Ascorbic Acid biosybthesis missing in man. *J Biol Chem*, 269: 13685-13688, 1994.

Plume: Promoting Economical, Useful and Maintained Software For The Higher Education And The Research Community

Christelle Dantec[1], Emmanuel Courcelle[2] and All contributors

[1] Montpellier GenomiX, UMR5203 CNRS, 141 rue de la Cardonille, 34094, Montpellier, Cedex 05, France
Christelle.dantec@igf.cnrs.fr.fr

[2] LIPM, UMR 2594/441 INRA-CNRS, BP52627, 31326, Castanet Tolosan, Cedex, France
Emmanuel.Courcelle@toulouse.inra.fr

Abstract *We describe here the project PLUME (FEATHER in the english version). This project references software used or developed in the Higher Education and Research community. It is built around a collaborative web site, each software is described by a regular user, or for « ESR » cards by the developer himself.*

Keywords Promote software, internal developments, listing, contribution.

Plume: Promouvoir les Logiciels Utiles, Maîtrisés et Economiques pour la communauté de l'Enseignement Supérieur et de la Recherche

Résumé *Nous décrivons ici le projet PLUME, qui référence les programmes utilisés ou développés dans la communauté Education-Recherche. Il est construit autour d'un site web collaboratif, chaque logiciel étant décrit par un utilisateur régulier, ou pour les fiches « ESR », par le développeur lui-même.*

Mots-clés Promouvoir les logiciels, développements internes, référencement, contribution.

1 Introduction

Le projet PLUME vise à **P**romouvoir les **L**ogiciels **U**tiles, **M**aîtrisés et **E**conomiques dans la communauté de l'**E**nseignement **S**upérieur et de la **R**echerche(ESR) en les référençant sur son site Web public http://www.projet-plume.org. « Utiles » signifie utilisés « Maîtrisés » veut dire que plusieurs sites utilisent ces logiciels couramment, tandis que « Économiques » regroupe au premier chef les logiciels libres, mais également les « gratuiciels » ou les logiciels diffusés sous licence propriétaire, mais avec des tarifs particulièrement avantageux pour l'ESR.

Plume a quatre objectifs principaux:

- Mutualiser les compétences sur les logiciels,
- Promouvoir les développements internes,
- Animer une communauté autour du logiciel,
- Promouvoir l'usage et la contribution aux logiciels libres.

Les contributions viennent des membres de cette communauté Enseignement Supérieur–Recherche, qui rédigent des fiches descriptives réparties par métiers et activité : Biologie, Chimie, Agronomie, Développeur, Calcul scientifique, Mathématiques, etc...

On trouve trois grandes catégories de fiches: des logiciels en production, des logiciels développés au sein de l'ESR, ainsi que des fiches « ressources ».

2 Contributions

2.1 Logiciels « validés »

Les « logiciels validés » référencés dans Plume sont des logiciels largement utilisés dans la communauté de l'Enseignement Supérieur et de la Recherche. Ce sont soit des logiciels généralistes, soit des logiciels « métiers » qui peuvent avoir une utilisation plus réduite du fait de l'étroitesse de leur champ applicatif. La fiche est relue avant publication par au moins deux personnes. Chaque logiciel est décrit par une fiche synthétique qui décrit les fonctionnalités, son opérabilité, sa pérennité et l'utilisation faite dans les laboratoires. Le contributeur met régulièrement à jour la fiche (environ tous les 6 mois) afin que le référencement suive les évolutions du lo-

giciel. Une fiche qui ne peut être mise à jour en raison de la défection de son contributeur est retirée du serveur.

2.2 Logiciels développés dans l' ESR

Ces fiches « ESR » décrivent les logiciels développés ou en cours de développement dans la communauté ESR Ces fiches sont destinées à valoriser les productions logicielles et à les faire connaître aux chercheurs/enseignants du domaine dans un but de collaboration scientifique. A chaque fiche sont associés les noms des développeurs, des laboratoires associés et des tutelles.

2.3 Ressources

Ces fiches présentent synthétiquement des ressources liées aux logiciels présentés dans PLUME et plus généralement aux logiciels libres ou utilisés dans la communauté ESR : articles, support de cours, événements, recommandations juridiques et administratives pour diffuser un logiciel de laboratoire, méthodologie pour tracer la propriété intellectuelle dans des codes logiciels, ...

3 Bioinformatique dans Plume

La bioinformatique se retouve dans le thème biologie http://www.projet-plume.org/biologie: Une cinquantaine de contributeurs ont rédigé actuellement plus de 30 fiches en rapport avec la biologie.

Le caractère « utile » et « maîtrisé » de ces logiciels est d'une part affirmé par les contributeurs, spécialistes du domaine et utilisateurs du logiciel, d'autre part par le fait qu'un logiciel n'est référencé dans Plume que s'il est en production dans au moins 3 sites différents. Cela permet d'avoir un avis pertinent en condition réelle d'utilisation du logiciel, et contribue à la valeur ajoutée de Plume par rapport à d'autres sites qui référencent aussi des logiciels utilisés en bioinformatique. De plus chaque fiche est reliée avec les logiciels connexes. Certains contributeurs acceptent de fournir un support léger aux logiciels décrits pour aider les lecteurs à L'installer.

De nombreux logiciels bioinformatiques sont développés dans les laboratoires et malheureusement restent méconnus. Plume permet de valoriser et diffuser ces développements.

Plume permet également par un système de recherche sur des mots clés de trouver rapidement un type de logiciel attendu. La bioinformatique regroupe des thèmes très variés et le travail de référencement des logiciels utiles et maitrisés en bioinfomatique est grand. Actuellement on trouve des fiches sur :

des applications de séquençage haut débit *(MACS, Galaxy, S-MART)*, le transcriptome *(BASE, GAGG)*,

les séquences *(Blast EMBoss, BioMaJ)*, la spectrométrie de masse *(massXpert)*, la chimie *(Zebre)*, l'agronomie *(Virtual Laboratory Environment)*, la génomique comparative *(Narcisse)*, les formats de fichiers *(.bar, .bed, .gff)*

Les formats de fichiers sont nombreux dans la discipline et actuellement peu sont encore référencés.

La plateforme Plume propose une organisation originale et bien structurée qui a pour but de maintenir ce service sur le long terme. Regroupant des logiciels issus de nombreux domaines scientifiques, la plateforme est appelée jouer un rôle important dans la communauté ESR française ou étrangère, de part sa notoriété. La communauté Bioinformatique a tout intérêt à utiliser et faire vivre ce service, afin de profiter de l'effet de synergie entre les disciplines apporté par la plateforme qui devrait permettre d'augmenter la visibilité des logiciels produits par cette communauté.

4 La communauté Plume

Toute personne de l'Enseignement Supérieur et de la Recherche peut participer à Plume en tant que membre (pour proposer des fiches logicielles ou commenter les fiches existantes) ou contributeur (afin de rédiger de nouvelles fiches) : On peut également proposer de s'investir de manière permanente en devenant « responsable thématique ».

A ce jour (Avril 2010), Plume regroupe 530 contributeurs (dont 55 pour le thème Biologie) et 1192 membres.

Développeurs et utilisateurs de logiciels « UME », nous attendons vos contributions: elles permettront de faire connaître à vos collègues les logiciels que vous appréciez. Dans le cas de fiches « ESR », en offrant un site centralisé de référence, PLUME peut vous aider à faire connaître plus largement vos productions ou celles de votre laboratoire, et ainsi contribuer à créer une communauté d'utilisateurs.

5 Remerciements

Ce projet, initié par Jean-Luc Archimbaud, est porté par le CNRS à travers l'UREC.

Merci aux centaines de contributeurs qui participent à ce projet.

The IntAct molecular interaction database and data distribution with PSICQUIC

Marine Dumousseau, Sandra Orchard, Bruno Aranda, Samuel Kerrien, Jyoti Khadake, Margaret Duesbury and Henning Hermjakob[1]

[1] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
hhe@ebi.ac.uk

Abstract *IntAct is an open-source, open data, molecular interaction database and toolkit. Data is abstracted from the literature or from direct data depositions by expert curators following a deep annotation model providing a high level of detail. As of June 2010, IntAct contains over 218,028 curated binary interaction evidences. The data are released on a weekly basis and are available in both XML (PSI-MI XML) and tabular format (PSIMITAB). IntAct is also a member of the International Molecular Exchange consortium (IMEx) that aims at sharing manual curation effort and exchange completed records between collaborating partners, thus increasing the amount of data made available to the scientific community. To be able to exchange and share molecular interaction annotations, IntAct implements the PSIQUIC services, making its data available to any PSIQUIC client. It also embeds a PSIQUIC client in its user interface, so that results from the IntAct interface are expanded by data from other databases such as MINT, BioGrid and DIP. PSICQUIC is an effort from the HUPO Proteomics Standard Initiative (HUPO-PSI) to standardise the access to molecular interaction databases programmatically. It is motivated by the idea that such annotations should not be provided by single centralized databases, but should instead be spread over multiple sites. Data distribution, performed by PSICQUIC servers, is separated from visualization, which is done by PSICQUIC clients. PSICQUIC is a client-server system in which a single client integrates information from multiple servers. It allows a single machine to gather up molecular interactions from multiple distant web sites, collate the information, and display it to the user in a single view.*

Keywords IntAct, PSICQUIC, molecular interaction, open-source, open data, HUPO-PSI, MIQL, standards, database

1 Introduction

IntAct is an open-source, open data, molecular interaction database and toolkit. Data is abstracted from the literature or from direct data depositions by expert curators following a deep annotation model providing a high level of detail from the experimental reports on the full text of the publication. IntAct aims to provide the user with all the relevant experimental detail described in the originating article, with all entries being fully compliant with the International Molecular Exchange consortium (IMEx - http://imex.sourceforge.net/) guidelines [1] and the Minimum Information required to report a Molecular Interaction Experiment (MIMIx- http://www.psidev.info/index.php?q=node/278) standard [2], whilst also providing extra levels of information beyond these minimum requirements.

12 The IntAct Molecular Interaction Database and its curation policy

IntAct makes extensive use of a number of controlled vocabularies, primarily the Molecular Interaction ontology of the Proteomics Standard Initiative (PSI-MI, [3]) to describe the technical details of the experiment, binding sites, protein tags and mutations. The Gene Ontology [4] is used to describe the sub-cellular location an interaction may be shown to occur in or the function of an enzyme in an enzyme/substrate assay. Interacting molecules are systematically mapped to stable identifiers from public databases such as UniProtKB for proteins [5], ChEBI [6] for small molecules, Ensembl [7] for genes and the DDBJ[8]/EMBL[9]/GenBank[10] nucleotide databases for nucleic acids. Binding sites are also cross-referenced to the InterPro database [11] whenever possible.

As of June 2010, IntAct contains over 218,028 curated binary interaction evidences.

Interactions from over 275 species are described, although our main focus is on human and model organisms such as mouse, Drosophila melanogaster, Caenorhabditis elegans, Escherichia coli and Arabidosis thaliana. In response to the growing data volume and user requests, IntAct provides a two-tiered view of the interaction data. A Quick Search box allows simple "Google-like" searches to be performed, however, the search interface allows the user to iteratively develop complex queries using the flexible Molecular Interaction Query Language (MIQL), exploiting the detailed annotation with hierarchical controlled vocabularies. Users can search for interactors that are annotated with terms from the Gene Ontology [12], InterPro, ChEBI and the UniProt Taxonomy, or interactions described using specific terms from the PSI-MI controlled vocabularies. Results are provided at any stage in a simplified, tabular view. Specialized views then allow 'zooming in' on the full annotation of interactions, interactors and their properties.

IntAct source code and data are freely available at http://www.ebi.ac.uk/intact. The data are released on a weekly basis and are available in both XML (PSI-MI XML) and tabular format (PSIMITAB).

13 IntAct and development of PSICQUIC, a common query interface

The Proteomics Standards Initiative Common Query InterfaCe (PSICQUIC) aims at standardizing the programmatic access to molecular interaction databases. On the simplest level, PSICQUIC provides a set of methods to simultaneously query multiple molecular interaction databases. The PSICQUIC interface allows single word or phrase queries (e.g. abl1 AND "pull down"), search in specific attributes/columns (e.g. abl1 AND species:human), wildcards (e.g. abl*), and logical operators, through the use of MIQL. PSICQUIC servers may return data in one or more output types, usually the tabular MITAB25 or the more detailed PSI-MI XML format. To be able to exchange and share molecular interaction annotations, IntAct actively participates in the development of PSICQUIC. IntAct implements the PSICQUIC service, making its data available to any PSICQUIC client. It also embeds a PSICQUIC client in its user interface, so that results from the IntAct interface are extended by data from other databases such as MINT, BioGrid and DIP. The ChEMBL database provides protein-chemical entity interactions, and Reactome provides simplified pathway information all through the same interface, enabling these to be merged into a single view from PSI-compliant tools such as Cytoscape. A full list of available services can be found on the PSICQUIC registry (www.ebi.ac.uk/Tool/webservices/PSICQUIC/registry/registry?action=STATUS).

14 PSICQUIC implementation

PSICQUIC is motivated by the idea that no one single centralized databases can provide all available interaction data, but by combining the data from multiple sites, the user benefits from increased coverage of the interactome of a particular organism. Data distribution, performed by PSICQUIC servers, is separated from visualization, which is done by PSICQUIC clients. PSICQUIC is a client-server system in which a single client can integrate information from multiple servers. It allows a single machine to gather up molecular interactions from multiple distant web sites, collate the information, and display it to the user in a single view. One immediate user of the PSICQUIC service has been the IMEx consortium. IMEx aims at sharing manual curation effort and providing a non-redundant set of high-quality, consistently annotated completed records provided by collaborating partners.

Acknowledgements

IntAct is supported by the European Commission under SLING, grant agreement number 226073 (Integrating Activity) within Research Infrastructures of the FP7, under PSIMEx, contract number FP7-HEALTH-2007-223411 and under APO-SYS, contract number FP7-HEALTH-2007-200767.

References

[1] Orchard S, Kerrien S, Jones P, Ceol A, Chatr-Aryamontri A, Salwinski L, Nerothin J, Hermjakob H. Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. (2007) 7(Suppl. 1):28–34

[2] Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stümpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat. Biotechnol. (2007) 25:894–898.

[3] Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, et al. Broadening the horizon—level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biol. (2007) 5:44.

[4] The Gene Ontology Project in 2008. Nucleic Acids Res. (2008) 36:D440–D444.

[5] The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res. (2009) 37:D169–D174

[6] Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. (2008) 36:D344–D350.

[7] Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, et al. Ensembl 2009. Nucleic acids research (2009) 37:D690–D697.

[8] Sugawara H, Ikeo K, Fukuchi S, Gojobori T, Tateno Y. DDBJ dealing with mass data produced by the second generation sequencer. Nucleic Acids Res. (2009) 37:D16–D18.

[9] Cochrane G, Akhtar R, Bonfield J, Bower L, Demiralp F, Faruque N, Gibson R, Hoad G, Hubbard T, Hunter C, et al. Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Res. (2009) 37:D19–D25.

[10] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. (2009) 37:D26–D31.

[11] Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. (2009) 37:D211–D215.

[12] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet (2000) 25:25–29.

IMGT/3Dstructure-DB and tools for immunoglobulins (IG) or antibodies, T cell receptors (TR), MHC, IgSF and MhcSF structural data

François EHRENMANN[1] and Marie-Paule LEFRANC[1]

[1] Laboratoire d'ImmunoGénétique Moléculaire, LIGM, Université Montpellier 2, UPR CNRS1142, 141, rue de la Cardonille, 34396, Montpellier, Cedex 5, France
{Francois.Ehrenmann,Marie-Paule.Lefranc}@igh.cnrs.fr

Keywords IMGT, immunoglobulin, antibody, T cell receptor, MHC, IgSF, MhcSF

1 Introduction

IMGT/3Dstructure-DB (1,2) is the three-dimensional (3D) structure database of IMGT®, the international ImMunoGenetics information system® http://www.imgt.org (3), the global reference in immunogenetics and immunoinformatics. Major breakthroughs characterize IMGT® and, therefore, IMGT/3Dstructure-DB: a standardized identification (IMGT keywords), a standardized nomenclature (IMGT gene and allele names), a standardized description (IMGT labels), and a standardized numerotation (IMGT unique numbering). IMGT-ONTOLOGY concepts have been crucial in bridging the gap between sequences and 3D structures (4,5) in IMGT/3Dstructure-DB database and in the IMGT/DomainGapAlign (2) and IMGT/Collier-de-Perles (6,7) tools.

2 IMGT/3Dstructure-DB

The IMGT/3Dstructure-DB structural data are extracted from the Protein Data Bank (PDB) and annotated according to the IMGT-ONTOLOGY concepts of classification, using internal tools and IMGT/DomainGapAlign. IMGT/3Dstructure-DB provides the closest genes and alleles that are expressed in the amino acid sequences of the 3D structures, by aligning these sequences with the IMGT domain reference directory. Each entry in the database is detailed in an IMGT/3Dstructure-DB card. Eight tabs are available at the top of each card: 'Chain details', 'Contact analysis', 'Paratope and epitope', '3D visualization Jmol or QuickPDB', 'Renumbered IMGT file', 'IMGT numbering comparison', 'References and links' and 'Printable card'. As an example, the 'Chain details' comprises information, first, on the chain itself (chain ID, chain length, IMGT chain description...), then on each domain, starting from the N-terminal end (IMGT domain description, gene and allele names...). IMGT/Colliers de Perles on two layers, available for the variable (V) and constant (C) type domains, are displayed with hydrogen bonds. 'Contact analysis' provides contacts between structural units (domains or ligand) and are obtained by a local program written in C in which atoms are considered to be in contact when no water molecule can take place between them. The atom contact types and categories for each amino acid are provided in 'IMGT Residue@Position cards'. 'IMGT pMHC contact sites' graphically represent, in IMGT Colliers de Perles, the MHC amino acid positions that contact the peptide side chains in pMHC complexes, and thus allow comparison of pMHC interactions. These features are part of the information provided by IMGT/3Dstructure-DB. In June 2010, the IMGT/3Dstructure-DB database manages 2242 coordinate files (PDB, INN and Kabat).

3 IMGT/Collier-de-Perles

IMGT Colliers de Perles are 2D graphical representations (6,7) available for the V type domain (V-DOMAIN of IG and TR, V-LIKE-DOMAIN of IgSF other than IG and TR), C type domain (C-DOMAIN of IG and TR, C-LIKE-DOMAIN of IgSF other than IG and TR) and groove (G) type domain (G-DOMAIN of MHC, G-LIKE-DOMAIN of MhcSF other than MHC). Any domain represented by an IMGT Collier de Perles is characterized by the length of its strands, loops and turns and, for the G type, by the length of its helix. IMGT Colliers de Perles are generated with the IMGT/Colliers-de-Perles tool which allows the users to draw IMGT Colliers de Perles starting from their own amino acid sequences. Sequences have to be gapped according to the IMGT unique numbering (using for example IMGT/DomainGapAlign). Adjustements are possible manually in the window for unusual features. IMGT/Collier-de-Perles tool can be customized to display the CDR-IMGT according to the IMGT Color menu or to visualize the amino acids according to their hydropathy, volume or IMGT physicochemical classes (8).

4 IMGT/DomainGapAlign

IMGT/DomainGapAlign is a tool which aligns the user amino acid sequences with the IMGT domain reference directory, identifies the closest V, C and G domains, creates gaps according to the IMGT unique numbering and highlights differences with the closest reference(s). For an antibody V domain sequence, IMGT/DomainGapAlign identifies the closest germline V-REGION and J-REGION, and provides a delimitation of the strands, framework regions (FR-IMGT) and CDR-IMGT. The IMGT gene and allele name of the closest sequence(s) from the IMGT domain reference directory is provided with a percentage of identity and a Smith-Waterman score. Regions and domains are highlighted using the IMGT Color menu. Several sequences can be analysed simultaneously and users can choose how many alignments to display for each sequence. The IMGT Colliers de Perles are generated from the gapped sequences provided by the IMGT/DomainGapAlign tool.

5 IMGT/DomainDisplay

IMGT/DomainDisplay is a tool which provides the display of amino acid sequences from the IMGT domain directory. This directory contains, for the IG and TR, amino acid sequences of the domains encoded by the C genes and the translation of the germline V and joining (J) genes. The identified genes are classified based on the IMGT nomenclature of IG and TR genes and alleles that was approved in 1999 by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC), entered in IMGT/GENE-DB, and endorsed by the World Health Organization (WHO)-International Union of Immunological Societies (IUIS) Nomenclature Committee. Entrez Gene at the National Center for Biotechnology Information (NCBI), Vertebrate Genome Annotation (Vega) at the Wellcome Trust Sanger Institute, and Ensembl at the European Bioinformatics Institute (EBI) currently use IMGT nomenclature.

6 Conclusion

IMGT/3Dstructure-DB and tools are widely used by researchers, particularly for antibody engineering and humanization design (1). Indeed they allow to precisely define and to easily compare amino acid sequences of the FR-IMGT and CDR-IMGT, between the nonhuman (mouse, rat…) and the closest human V domains. A recent analysis performed on humanized antibodies used in oncology underlines the importance of a correct delimitation of the CDR to be grafted. IMGT/3Dstructure-DB facilitates the identification of potential immunogenic residues at given positions in chimeric or humanized antibodies, including those of the C domains. These therapeutic applications emphasize the importance of the IMGT/3Dstructure-DB standardized approach that bridges the gap between sequences and 3D structures whatever the species. Since 2008, amino acid sequences of monoclonal antibodies (mAb, suffix -mab) and of fusion proteins for immune applications (FPIA, suffix -cept) from the WHO/International Nonproprietary Name (INN) programme have been entered in IMGT®.

References

[1] Q. Kaas, M. Ruiz and M.-P. Lefranc, IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res*., 32, D208-D210, 2004.

[2] F. Ehrenmann, Q. Kaas and M.-P. Lefranc, IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res.* 38(Database issue):D301-7, 2010.

[3] M.-P. Lefranc, V. Giudicelli, C. Ginestoux, J. Jabado-Michaloud, G. Folch, F. Bellahcene, Y. Wu, E. Gemrot, X. Brochet, J. Lane, L. Regnier, F.Ehrenmann, G. Lefranc and P. Duroux, IMGT®, the international ImMunoGeneTics information system®. Nucleic Acids Res., 37,1006-1012, 2009.

[4] Q. Kaas, F. Ehrenmann and M.-P. Lefranc, IG, TR and IgSF, MHC and MhcSF: what do we learn from the IMGT Colliers de Perles? Brief. Funct. Genomic Proteomic, 6:253-264, 2007.

[5] M.-P. Lefranc, V. Giudicelli, L. Regnier and P. Duroux, IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform*., **9**, 263-275, 2008.

[6] M. Ruiz and M.-P. Lefranc, IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics*, **53**, 857-883, 2002.

[7] Q. Kaas and M.-P. Lefranc, IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. *Current Bioinformatics*, **2**, 21-30, 2007.

[8] C. Pommié, S. Levadoux, R. Sabatier, G. Lefranc and M.-P. Lefranc IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties *J. Mol. Recognit.*, **17**, 17-32, 2004.

MetaboFlux: a method to analyse flux distribution in metabolic networks

Amine GHOZLANE[12], Frédéric BRINGAUD[3], Fabien JOURDAN[4] and Patricia THÉBAULT[12]

[1] LaBRI, UMR 5800 CNRS, Université Bordeaux 1, 351, cours de la Libération, 33405 Talence, France
amine.ghozlane@labri.fr
[2] CBiB, Université Bordeaux 2, 142 rue Léo Saignat, 33076 Bordeaux, France
patricia.thebault@u-bordeaux2.fr
[3] RMSB, UMR 5536 CNRS, Université Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux, France
frederic.bringaud@rmsb.u-bordeaux2.fr
[4] INRA, UMR 1089 Xénobiotiques, 180 Chemin de Tournefeuille, 31027 Toulouse, France
fabien.jourdan@toulouse.inra.fr

Abstract *Robust reconstruction of metabolic networks has become an important challenge in biology and the definition of powerfull methods for model validation constitute a critical step. To this end, we propose a new modelling approach that helps to investigate the system's phenotype and its semi-quantitative behavior by accrying out metabolic flux analysis. We have implemented this method in "Metaboflux", and applied it to the Trypanosoma brucei energetic metabolism to determine fluxes in each individual glycosomal branches compatible with biological observations.*

Keywords Petri Nets, Flux analysis, Metabolic network, *Trypanosoma brucei*.

1 Introduction

The main objective of this paper, given a model and some experimental data, is to propose a method for the prediction of fluxes and model validation (identified as a critical part [1]) in metabolic networks. Flux analysis and model validation may be both addressed by carrying out Petri Net (PN) simulations to study the behavior of a model and confront it to the available biological data.

Flux balance analysis (FBA) gives a mathematical framework [2,3] to analyze metabolic capabilities of an organism in a constraint-based model. It calculates all of the feasible chemically balanced metabolic routes through the network by maximizing an unique objective function (biomass production, a metabolite concentration). Despite the success of FBA and several context specific extensions (as MOMA[4], rFBA[5]), the definition of the objective function remains a challenge for improving the biological meaning [6]. This task may be yield by integrating additional constraints for increasing biological assumptions in a multi-objective function. In such analysis, the complexity of the underlying problems involves the use of meta-heuristics.

We present a new method for running fluxes analyses when integrating heterogeneous available data. The structural complexity of metabolic networks can lead to alternatives fluxes that can be formulated as a complex optimization problem. Most of the real values of the fluxes are unknown and the biologically realistic range for most parameters may span large intervals of values. We have applied a heuristic algorithm to compute optimal flus distribution solutions. The key question that our method addresses is to identify any set of parameter values for which the network model would exhibit a realistic behavior, given initial conditions. Furthermore, we propose several scenarii of fluxes for biological expertise as the developped heuristic may propose a set of closed optimal solutions.

2 Method

The known biological informations are modelled by a stochastic PN (transitions are given for the reactions and places for metabolites) where delays can be assigned to transitions as a probability distribution. This variant of PN allows the specification of flux ratio that will be tested through the simulation process. From a given set of probability distributions representing the flux amount of reactions (the input set of parameters), the simulation of the PN allows the exploration of all possible behaviors of the system. At the end of a run, if all input metabolites are consumed, we get the concentrations of the intermediate and output metabolites. We integrate the expected metabolites concentrations and/or some known fluxes revealed by biological experiments (when available) within a multiple objective function (Eq.1 on the following page) that has to be optimized. This function defines the biological con-

straints that have to be satisfied for getting a realistic model. The energy value is calculated by the objective function where $X_{i_{init}}$ stands for the known metabolite concentration or for the known flux ratio (when they are available), $X_{i\,final}$ stands for the corresponding values resulting from a PN simulation and w_i as a normalizing factor.

$$Energy = \sqrt{\sum_{i=0}^{n} (w_i * (\frac{X_{i_{final}} - X_{i_{init}}}{X_{i_{init}}})^2)} \quad (1)$$

During this stage, the simulation process (resulting from the Petri net model) is encapsulated in a simulated annealing process [7] that ends by a Nelder-Mead minimization stage [8]. Simulations are carried out by fitting the set of input parameters until the system reaches the best optimization given by the lowest Energy value. To explore a large set of possible behaviors of the system, several runs of simulations are interlaced with the simulated annealing process (Fig. 1). A set of solutions is given by the best cluster of flux distributions, that best fits the expected metabolites concentrations. The process is repeated giving several scenarios for biological expertise. Our methodological development are implemented in the standalone MetaboFlux software (http://www.cbib.u-bordeaux2.fr/metaboflux).

3 Case study

MetaboFlux has been applied to study the glucose metabolic network of a parasitic protist of vertebrates that causes sleeping sickness in Africa, *Trypanosoma brucei*. A major part of glucose metabolism of the procyclic form of *T. brucei*, including the 6 first glycolytic steps, occurs in an organelle called glycosome. The glucose metabolic network, including the glycosomal contribution, has been built by exploiting genomic, reverse genetic and metabolomic data [9]. Some known biological constraints, such as

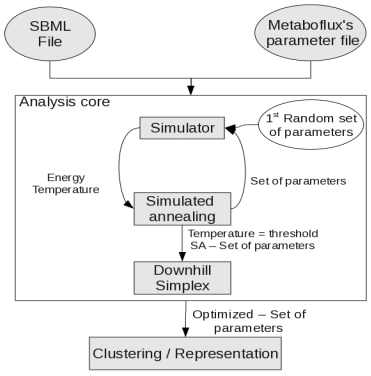


Fig. 1. The pipeline of MetaboFlux. The network structure is given via a SBML file.

the maintenance of the glycosomal ATP/ADP and $NADH/NAD^+$ balances, have not been carefully addressed in the current model. Indeed, glycosomes are peroxisomes-like organelles, which are not supposed to exchange ATP/ADP and $NADH/NAD^+$ molecules with the cytosol. Consequently, metabolic fluxes in the different branches of the metabolic network have to be compatible with the maintenance of these balances. The resulting scenarios, given by MetaboFlux, strongly support the current metabolic model. This analysis determined metabolism fluxes in each individual branches compatible with known constraints. Metaboflux was designed in such a way that biologists will be able to test the model with new data, to define new relevant fluxes scenarios.

4 Conclusion

In this work, a novel method was presented that allows to analyse the structural complexity of metabolic networks by running efficient dynamic simulations for fluxes prediction. MetaboFlux will be able to assist biologists to test different structural models by calculating their optimal metabolic behaviors according to available heteregoneous biological data and shows interesting performances for the model validation task.

Acknowledgements

This work is supported by the Agence Nationale de la recherche (ANR) program METABOTRYP of the ANR-MIME2007 and SYSTRYP of the ANR-BBRCS call and the CBiB.

References

[1] Wiechert W. *Modeling and simulation: tools for metabolic engineering.* J Biotechnol 94:37-63, 2002
[2] D.A. Fell, and J.R. Small, *Fat synthesis in adipose tissue. an examination of stoichiometric constraints*, Biochem J., 238(3):781-786, 1986.
[3] Stelling J. *Mathematical models in microbial systems biology.* Curr Opin Microbiol 7:513-518, 2004
[4] Segrè D, Vitkup D, Church GM. *Analysis of optimality in natural and perturbed metabolic networks.* Proc Natl Acad Sci U S A 99:15112-15117, 2002
[5] Covert M, Schilling C, Famili I, Edwards J, Goryanin I, Selkov E, Palsson B. *Metabolic modeling of microbial strains in silico* Trends in Bioch Sc 26:179-186, 2001
[6] Gianchandani EP, Chavali AK, Papin JA. *The application of flux balance analysis in systems biology*, Wiley Interdis Rev: Systems Biol and Med 2:372-382, 2010
[7] Kirkpatrick S, Gelatt CD, Vecchi MP. *Optimization by Simulated Annealing.* Science 220:671-680, 1983
[8] Nelder JA, Mead R. *A Simplex Method for Function Minimization* The Computer Journal 7:308-313, 1965
[9] Bringaud F, Rivière L, Coustou V . *Energy metabolism of trypanosomatids: adaptation to available carbon sources.* Mol Biochem Parasitol 149:1-9, 2006

IMGT-ONTOLOGY for immunogenetics and immunoinformatics information systems

Véronique GIUDICELLI[1] and Marie-Paule LEFRANC[1]

[1] Laboratoire d'ImmunoGénétique Moléculaire, LIGM, Université Montpellier 2, UPR CNRS 1142, Institut de Génétique Humaine, IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, Montpellier, France
{Véronique.Giudicelli, Marie-Paule.Lefranc}@igh.cnrs.fr

Abstract *IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics, manages the immunogenetics knowledge through diverse facets relying on seven axioms and represents a paradigm for the elaboration of integrated ontologies in system biology.*

Keywords IMGT, immunogenetics, immunoinformatics, ontology, information system, system biology.

1 Introduction

IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org), is the reference in immunogenetics and immunoinformatics. IMGT® standardizes and manages the complex immunogenetics data which include the immunoglobulins (IG) or antibodies, the T cell receptors (TR), the major histocompatibility complex (MHC) and the related proteins of the immune system (RPI) which belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF) [1]. The accuracy and consistency of IMGT® data and the coherence between the different IMGT® components (databases, tools and Web resources) are based on IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics [2]. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets relying on seven axioms, IDENTIFICATION, CLASSIFICATION, DESCRIPTION, NUMEROTATION, LOCALIZATION, ORIENTATION and OBTENTION, that postulate that objects, processes and relations have to be identified, described, classified, numerotated, localized, orientated, and the way they are obtained, determined. These axioms constitute the Formal IMGT-ONTOLOGY, also designated as IMGT-Kaleidoscope [3].

2 IMGT-ONTOLOGY axioms and concepts

The IDENTIFICATION axiom has generated the concepts of identification which allow to identify any biological objects, processes and relations in IMGT®. They provide the terms and rules that were necessary to define the IMGT standardized keywords used, in IMGT® databases, for the identification of IG, TR or MHC nucleotide and protein sequences, and structures according to their fundamental biological and immunogenetics characteristics.

The CLASSIFICATION axiom provides the rules that are necessary to classify the IG and TR genes. Indeed, the IG and TR genes belong to highly polymorphic multigene families organized as clusters in several loci in the genome [4,5]: therefore their classification requires a strong knowledge standardization. As a major contribution, the concepts of classification allowed to set up a unique nomenclature of human IG and TR genes, which was approved by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 and has become the community standard. The IG and TR genes are managed in the IMGT/GENE-DB database [6]. IMGT/GENE-DB database entries are cross-referenced by HGNC database, Genatlas, Entrez Gene (NCBI) and Vega (Wellcome Trust Sanger Institute).

The DESCRIPTION axiom and related concepts correspond to the standardization of terms and rules which allow to describe the structural and functional characteristic features of the IG, TR and MHC nucleotide and protein sequences, and 3D structures. Description concepts include IMGT standardized labels and the topological relationships that are used for the annotation process. Interestingly, 64 of the IMGT labels have been integrated by Sequence Ontology (http://www.sequenceontology.org/).

The NUMEROTATION axiom and the concepts of numerotation determine the principles of a unique numbering for variable, constant and groove domains in IG, TR and MHC sequences and 3D structures. The concept of 'IMGT unique numbering' and its graphical representation, the 'IMGT Collier

de Perles' represent a major breakthrough and are the flaship of IMGT® since they allow to bridge the gap between sequences and structures [7,8]. Thanks to both concepts, IMGT® provides a standardized and a fine description of allelic polymorphisms of IG and TR genes and of the somatic mutations in IG sequences. It also provides the rules for the delimitations of the framework and complementarity determining regions of IG and TR allowing the standardization of the contact analysis between residues in 3 D structures.

The LOCALIZATION axiom postulates that molecules, cells, organs, organisms or populations and their processes and relations have to be localized in time or space. At the molecular level in the field of immunogenetics, the concepts of localization allow to characterize the localization of IG, TR and MHC genes and proteins.

The ORIENTATION axiom defines the rules for the orientation of objects in IMGT®. In the context of genome analysis, it has led to set the 'Genomic orientation' concept (for chromosome, locus and gene) and the 'DNA strand orientation' concept.

The OBTENTION axiom has generated a set of standardized terms that precise, for any object of IMGT®, its origins (‘Origin’ concept) and the conditions in which the sequences have been obtained (‘Methodology’ concept).

3 Conclusion

The axioms of IMGT-ONTOLOGY have been essential for the conceptualization of the molecular immunogenetics knowledge and for the creation of IMGT®. All the components of the IMGT® integrated system have been developed, based on standardized concepts and relations, making IMGT® a system and an ontology that bridge biological and computational spheres in bioinformatics [9]. IMGT-ONTOLOGY concepts are available, for the biologists and IMGT® users, in the IMGT Scientific chart [1]. They are being formalized step by step in OWL language with the Protégé ontology editor. A first version has been published on NCBO Bioportal site (http://bioportal.bioontology.org/) and is also available from the 'IMGT downloads'. The concepts of IMGT-ONTOLOGY are currently used for the exchange and the sharing of knowledge in: (i) fundamental and medical research (repertoire analysis of the IG antibody sites and of the TR recognition sites in normal and pathological situations such as autoimmune diseases and infectious diseases), (ii) veterinary research, (iii) genome diversity and genome evolution studies of the adaptive immune responses, (iv) structural evolution of the IgSF and MhcSF proteins, (v) biotechnology related to antibody engineering (scFv, phage displays, combinatorial libraries, chimeric, humanized and human antibodies), (vi) diagnostics and (vii) therapeutical approaches (grafts, immunotherapy, vaccinology). IMGT-ONTOLOGY was a key component in the elaboration and setting up of standards of the European ImmunoGrid project (http://www.immunogrid.org/) whose aim is to define the essential concepts for modelling of the immune system. IMGT-ONTOLOGY can also be used for multi-scale level approaches at the molecule, cell, tissue, organ, organism or population level, emphasizing the generalization of the application domain. In that way IMGT-ONTOLOGY represents a paradigm for the elaboration of ontologies for immunogenetics and immunoinformatics information systems.

References

[1] M.-P. Lefranc, V. Giudicelli, C. Ginestoux, J. Jabado-Michaloud, G. Folch, F. Bellahcene, Y. Wu, E. Gemrot, X. Brochet, J. Lane, L. Regnier, F.Ehrenmann, G. Lefranc and P. Duroux, IMGT®, the international ImMunoGeneTics information system®. Nucleic Acids Res., 37,1006-1012, 2009.

[2] V. Giudicelli and M.-P. Lefranc, Ontology for Immunogenetics: IMGT-ONTOLOGY. Bioinformatics, 15:1047-1054, 1999.

[3] P. Duroux, Q. Kaas, X. Brochet, J. Lane, C. Ginestoux, M.-P. Lefranc and V. Giudicelli, IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. Biochimie, 90:570-583, 2008.

[4] M.-P. Lefranc and G. Lefranc, The Immunoglobulin FactsBook, Academic Press, London, UK, 2001.

[5] M.-P. Lefranc and G. Lefranc, The T Cell Receptor FactsBook, Academic Press, London, UK, 2001.

[6] V. Giudicelli, D. Chaume and M.-P. Lefranc, IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res., 33:256-261, 2005.

[7] M.-P. Lefranc, C. Pommié, R. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet and G. Lefranc, IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev. Comp. Immunol., 27:55-77, 2003.

[8] Q. Kaas, F. Ehrenmann and M.-P. Lefranc, IG, TR and IgSF, MHC and MhcSF: what do we learn from the IMGT Colliers de Perles? Brief. Funct. Genomic Proteomic, 6:253-264, 2007.

[9] M.-P Lefranc, V. Giudicelli, L. Regnier and P. Duroux, IMGT®, a system and an ontology that bridge biological and computational spheres in bioinformatics. Brief. Bioinform., 9:263-275, 2008.

CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources

David GOUDENEGE, Stéphane AVNER, Céline LUCCHETTI-MIGANEH, Frédérique BARLOY-HUBLER

LABORATOIRE, UMR6026 CNRS, SP@RTE, Bat13 campus Beaulieu, Université Rennes 1, 35000 Rennes, France
{david.goudenege, stephane.avner, celine.lucchetti, fhubler}@univ-rennes1.fr

Abstract *The functions of proteins are strongly related to their localization in cell compartments (cytoplasm or membranes) but the experimental determination of the sub-cellular localization of proteomes is laborious and expensive. A fast and low-cost alternative approach is in silico prediction, based on features of the protein primary sequences. However, biologists are confronted with a very large number of computational tools that use different methods that address various localization features with diverse specificities and sensitivities. As a result, exploiting these computer resources to predict protein localization accurately involves querying all tools and comparing every prediction output; this is a painstaking task. Therefore, we developed a comprehensive database, called CoBaltDB, that gathers all prediction outputs concerning complete prokaryotic proteomes. The current version of CoBaltDB integrates the results of 43 localization predictors for **784** complete bacterial and archaeal proteomes. CoBaltDB supplies a simple user-friendly interface for retrieving and exploring relevant information about predicted features (signal peptide, transmembrane segments). Data are organized into three work-sets (”specialized tools“, “meta-tools“ and “additional tools“). The database can be queried using the organism name or a list of locus tags and may be browsed using numerous graphical and text displays. With its new functionalities, CoBaltDB is a novel powerful platform that provides easy access to the results of multiple localization tools and support for predicting prokaryotic protein localizations with higher confidence than previously possible. CoBaltDB is available at:*
http://www.umr6026.univ-rennes1.fr/english/home/research/basic/software/cobalten

Keywords Tools database, subcellular localization, prokaryotic annotation.

1 Introduction

Determining the subcellular localization of proteins is essential for the functional annotation of proteomes [1]. Bacterial proteins can exist in soluble (i.e free) forms in cellular spaces (cytoplasm and periplasm in diderms), anchored to membranes or cell wall (in monoderms). They can also be released into the extracellular environment or directly translocated into host cells [3]. All protein synthesis takes place in the cytoplasm, so all non-cytoplasmic proteins must pass through one or two lipid bilayers by a mechanism called “secretion”.

Establishing whole proteome subcellular localization by biochemical experiments is possible but arduous, time consuming and expensive. Data concerning predicted proteins (from whole genome sequences) is continuously increasing. High-throughput in silico analysis is required for fast and accurate prediction of additional attributes based solely on their amino acid sequences. There are large numbers of global (that yield final localization) and specialized (that predict features) tools for computer-assisted prediction of protein localizations, using various methods.

This plethora of protein localization predictors and databases constitutes an important resource but requires time and expertise for efficient exploitation. Some of the tools require computing skills, as they have to be locally installed; others are difficult to use (numerous parameters) or to interpret (large quantities of graphics and output data). Web tools are disseminated and need numerous manual requests. Additionally, researchers have to decide which of these numerous tools are the most pertinent for their purposes, and selection is problematic without appropriate training sets. Recent work shows that the best strategy for exploiting the various tools is to compare them [3,4]. Here, we describe CoBaltDB [5], the first public database that displays the results obtained by 43 localization predictor tools for 776 complete prokaryotic proteomes.

2 Construction and Content

The CoBaltDB database contains three main types of data: i) prediction using 23 feature-based localization tools, ii) prediction obtained using 5 localization meta-tools and iii) data collected from 20 public database. These data were organized in five "boxes" with regard to the features predicted: three boxes correspond to signal peptide detection (Lipoprotein, Tat- and Sec- dependent targeting signals); one box for the prediction of alpha-transmembrane segments (TM-Box); and one box for outer membrane beta-barrels prediction. We retrieved and tested 99 currently (in 2009) available specialized and global tools that use various amino acid features and diverse methods (HMM, NN, SVM...). Some tools are Gram specific, for these reasons we have sorted the genome by phylogeny [2]. Currently, CoBaltDB contains pre-computed results obtained with 48 tools and databases, and additionally provides pre-filled access to 50 publicly available tools. Web-based tools were requested via a Web automat (http request) and standalone tools were installed on a Unix platform. The global python pipeline used multithreading to speed up the precomputation. The CoBaltDB platform has been developed as a Java client-server application. The server is installed at the Genouest Bioinformatics platform. An applet version is envisaged.

3 Utility

Our goal was to build an open-access reference database providing access to protein localization predictions. CoBaltDB was designed to centralize different types of data and to interface them so as to help researchers rapidly analyse and develop hypotheses concerning the subcellular distribution of particular protein(s) or a given proteome.

It presents four tabs that perform specific tasks: the "input" tab allows selecting the organism or a list of locus tags. The "Specialized tools" tab supplies a table showing, for each protein some annotation information and for each feature box (Tat, Sec, Lipo, aTMB, bBarrel), a heat map representing the percentage of tools predicting the truth/presence of the corresponding localization feature. Clicking on the heat map opens a new window that shows the tools raw data. The proteins which are referred as having an experimentally determined localization appear in yellow. The "meta-tools" tab provides the predictions given by meta-tools and global databases. The "additional tools" tab enables queries to be submitted to a set of 50 additional tools. Finally, for each protein, all results were summarized in a synopsis; the synopsis presents the results generated by all the tools in a unified manner, and includes a summary of all predicted cleavage sites and membrane domains. In order to allow the investigators to establish their own hypotheses and conclusions.

4 Conclusion

We have developed CoBaltDB, the first friendly interfaced database that compiles a large number of in silico subcellular predictions concerning whole prokaryotic proteomes. Currently, CoBaltDB allows fast access to precomputed localizations for 2,548,292 proteins in 784 proteomes. In all our analyses with CoBaltDB, it became clear that that the combination and comparative analysis of results of heterogeneous tools improved the computational predictions, and contributed to identifying the limitations of each tool. Therefore, CoBaltDB can serve as a reference resource to facilitate interpretation of results and to provide a benchmark for accurate and effective in silico predictions of the subcellular localization of proteins. Users can easily create small datasets and determine their own thresholds for each predicted feature (type I or II SPs for example) or proteome.

Acknowledgements

This work is supported by the Ministère de la Recherche. Thank to the Biogenouest platform.

References

[1] Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y: Automatic prediction of protein function. *Cell Mol Life Sci* 2003, 60(12):2637-2650.

[2] Desvaux M, Hebraud M, Talon R, Henderson IR: Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends in microbiology* 2009, 17(4):139-145.

[3] Park S, Yang JS, Jang SK, Kim S: Construction of Functional Interaction Networks through Consensus Localization Predictions of the Human Proteome. J Proteome Res 2009, 8(7):3367-3376.

[4] Shen YQ, Burger G: 'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools. BMC Bioinformatics 2007, 8:420.

[5] Goudenège D, Avner S, Lucchetti-Miganeh C, Barloy-Hubler F: CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources. BMC Microbiol. 2010 Mar 23;10:88.

Origin of Phenotypic Specificities in Wine Yeast through a Genomic Approach

Cyprien GUÉRIN, Hélène CHIAPELLO and Pierre NICOLAS

Unité Mathématique, Informatique & Génome, UR1077 INRA, Domaine de Vilvert, 78352, Jouy-en-Josas, France
{cyprien.guerin, helene.chiapello, pierre.nicolas}@jouy.inra.fr

Keywords *Saccharomyces cerevisiae*, wine yeast, phenotype, polymorphism.

Origine des Spécificités Phénotypiques de la Levure Œnologique a travers une Approche Génomique

Mots-clefs *Saccharomyces cerevisiae*, levure œnologique, phénotype, polymorphisme.

1 Introduction

La levure *Saccharomyces cerevisiae* est un champignon unicellulaire microscopique présent naturellement dans différentes niches écologiques. Elle est utilisée depuis l'antiquité dans la transformation de nombreux aliments, et en particulier en vinification.

Il existe de fortes variations de phénotypes entre les différentes souches de *S. cerevisiae*. La souche *EC1118* de levure œnologique industrielle a été sélectionnée pour ses caractéristiques d'initiation de la fermentation alcoolique en production vinicole [1]. Elle possède ainsi des propriétés phénotypiques spécifiques qui la différencient de la souche modèle de laboratoire *S288C* séquencée en 1996 [2]. La compréhension de l'origine génétique des propriétés fermentaires spécifiques des levures œnologiques est un enjeu majeur des prochaines années.

Les propriétés spécifiques d'une spore haploïde, *59A*, issue de la souche diploïde œnologique industrielle *EC1118* sont maintenant bien documentées mais l'origine génétique de ses caractéristiques phénotypiques n'est pas encore connue.

Une recherche de *Quantitative Trait Loci* (*QTL*) [3] dans une population de ségrégants, issue d'un croisement entre *S288C* et *59A* a permis de mettre en évidence deux régions influençant les caractéristiques phénotypiques d'intérêt, liés à la fermentation, et les niveaux de transcriptions d'un grand nombre de gènes (*eQTL*). Ces deux régions d'environ 30 000 et 76 000 paires de bases contiennent respectivement 16 et 32 gènes (données non publiées).

Dans ce travail, nous proposons : (i) d'analyser les polymorphismes existants entre le génome de la souche *S288C* et celui de la spore *59A* puis (ii) une analyse permettant de proposer une liste de gènes candidats dont les polymorphismes pourraient être à l'origine des principaux phénotypes d'intérêt fermentaires.

2 Polymorphismes entre *S288C* et *59A*

Pour comparer le génome de la souche *59A* à celui, déjà connu, de *S288C* (SGD : 8 avril 2008) [4,5], un séquençage de type Illumina/Solexa [6] a été réalisé (8 millions de lectures appariées de deux fois 35 nucléotides). Les différences entre les deux génomes ont été identifiées en alignant les lectures sur le génome de référence avec la suite logicielle MAQ (version 0.7.1) [7], spécialisée dans l'assemblage de données de séquençage haut-débit prenant en compte leur qualité. L'assemblage a été effectué en autorisant 3 différences nucléotidiques par lecture au maximum.

Cette méthode a permis de détecter 46 592 différences nucléotidiques entre les génomes de la référence *S288C* et celui de la spore *59A*, dont 29 270 concernent des régions annotés comme codantes (CDS) et 11 000 entrainent des changements protéiques. Le pourcentage d'éléments fonctionnels modifiés par des polymorphismes sont décrit dans le tableau ci-dessous (voir Tab. 1). On remarque la proportion importante de séquences protéiques modifiées par au moins un polymorphisme (60%).

3 Analyse Populationnelle de l'Effet des Polymorphismes sur les Phénotypes

Les deux régions d'intérêt influençant les caractéristiques phénotypiques d'intérêt fermentaires (*QTL* et *eQTL*) contiennent respectivement 5 et 15 gènes codant pour des protéines polymorphes. Il est donc nécessaire de cibler, parmi tous ces gènes, des

Nom des éléments fonctionnels	Pourcentage d'éléments fonctionnels modifiés (Nbr modif / Nbr total)	
Gène	76,22 %	(5 012 / 6 576)
Pseudogène	57.14 %	(12 / 21)
ANRt	8,73 %	(24 / 275)
ARNr	4,00 %	(1 / 25)
Site de régulation	1,77 %	(54 / 3 057)
Protéines	*60,48 %*	*(3 977 / 6 576)*

Tab. 1. Proportion des principaux éléments fonctionnels du génome de la levure [5] modifiés par des polymorphismes nucléotidiques existants entre S288C et 59A.

candidats dont on évaluera expérimentalement l'implication dans les variations phénotypiques. Pour cela, on s'est appuyé sur des données de séquence de 32 autres souches représentatives de la diversité des *S. cerevisiae* d'origines diverses, disponibles publiquement au *Sanger Institute* [8]. Toutes ces données de séquence correspondent à des spores haploïdes, représentant 168 136 nouvelles positions polymorphes. Pour ces 32 souches et pour S288C et 59A, trois valeurs de phénotypes ont été mesurées en phase initiale de fermentation (S. Dequin, communication personnelle) :

(1) V_{max} : vitesse maximum de fermentation
(2) V_{50} : vitesse à mi-fermentation
(3) Tf : temps de fermentation nécessaire pour produire 79g de CO_2

Les *QTL* associent les phénotypes de la V_{max} à la première région d'intérêt et de la V_{50} et du Tf à la deuxième.

À partir de ces données, nous avons utilisé un modèle linéaire pour décomposer le phénotype y d'une souche donnée s comme la somme d'un phénotype moyen μ, d'un effet γ du genotype de cette souche $g(s)$ et d'un bruit $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

$$y_s = \mu + \gamma_{g(s)} + \varepsilon$$

On teste ainsi l'hypothèse nulle selon laquelle il n'y a pas d'effet du génotype des souches étudiées sur les variations du phénotype considéré :

$$\begin{cases} H_0 : \gamma_{g(s)} = 0 \text{ , } \forall g(s) \\ H_1 : \exists\, g(s) \neq g(s') \text{ tel que } \gamma_{g(s)} \neq \gamma_{g(s')} \end{cases}$$

Une analyse de variance (ANOVA), effectuée pour chacun des trois phénotypes, permet de quantifier le rejet de H_0. La *p-value* obtenue donne donc une indication de la significativité de la corrélation phénotype-génotype pour chaque site (acide aminé) polymorphe, entre *S288C* et *59A*, au sein des gènes des deux régions d'intérêt fermentaire.

Nous avons ainsi mis en évidence 5 sites polymorphes dans la première région, sur 18 total, dont la repartition des génotypes est corrélée à la V_{max} (avec une même *p-value* $= 0,013$). Un gène de cette région porte 4 de ces sites, ce qui peut-être expliqué par un déséquilibre de liaison dû à leur proximité physique.

Parmi les 43 sites polymorphes de la deuxième région, 3 sites de 2 gènes différents sont corrélés au Tf (*p-value* $= 0,019$ pour 2 des sites et *p-value* $= 0,002$ pour l'autre) et 4 sites de 4 gènes différents sont corrélés à la $V50$ (*p-values* $\leq 0,048$).

Après évaluation du taux de faux positifs par permutation, les niveaux atteints ne sont pas suffisant pour confirmer l'existence de l'effet d'un des polymorphismes sur les phénotypes associés à la fermentation (avec un seuil à 5 %). Par contre, les gènes associés aux *p-values* les plus faibles sont des candidats privilégiés sur lesquels effectuer un test expérimental d'impact fonctionnel.

4 Conclusion

Le gène candidat de la première région, portant le plus de sites corrélés, a été validé expérimentalement comme étant à l'origine de la différence de vitesse maximum de fermentation existant entre la souche de laboratoire et l'haploïde œnologique [9].

Pour la deuxième region, dont les phénotypes associés de Tf et de V_{50} sont correlés, les deux gènes mis en évidence ici pour leur implication dans ces deux phénotypes sont actuellement en cours de validation expérimentale.

En conclusion, l'approche proposée permet une sélection, sans *a priori* d'annotation fonctionnelle, des gènes candidats potentiellement à l'origine des spécificité phénotypiques de la levure œnologique.

Remerciements

Ce travail est financé par le projet GENYEAS-TRAIT (ANR Blanc 2007-2010) et a été effectué en collaboration avec l'UMR Science Pour l'Œnologie de l'INRA de Montpellier (B. Blondin, S. Dequin).

Références

[1] M. Novoa *et al.*, PNAS, 2009.
[2] A. Goffeau *et al.*, Science, 1996.
[3] RB. Brem *et al.*, Science, 2002.
[4] JM. Cherry *et al.*, Nucleic acids research, 1998.
[5] http ://www.yeastgenome.org
[6] http ://www.illumina.com/
[7] H. Li *et al.*, Genome Research, 2008.
[8] G. Liti *et al.*, Nature, 2009.
[9] C. Ambroset, Thèse Montpellier Supagro, 2009.

Structural Analysis of Proteins with Tandem Repeats by Hybrid Approaches

Andrey V. KAJAVA

CENTRE DE RECHERCHES DE BIOCHIMIE MACROMOLECULAIRE
UMR 5237 CNRS, 1919 Route de Mende, 34293 Montpellier, Cedex 5, France
Andrey.Kajava@crbm.cnrs.fr

Abstract *A significant portion of proteins carrying fundamental functions contains arrays of tandem repeats. This presentation will provide a survey of current challenges related to these proteins including: identification of repeats in proteomes, prediction and molecular modeling of their 3D structures, applications of these knowledge to the annotation of genomes. Examples of successful application of hybrid approaches for prediction of the 3D structures of these proteins that combine bioinformatics analysis, molecular modeling, x-ray fiber diffraction, electron microscopy, STEM mass measurements, optical spectroscopy, and other biophysical techniques will be presented.*

Keywords Structural bioinformatics, repeats , amino acid sequence, protein structure .

Genome sequencing projects are revealing a large number of biologically important proteins having the tandem arrays of up to 40 residue repeats [1, 2]. A significant portion of these proteins carry fundamental biological functions. Furthermore, over the last years a number of evidences has been accumulated about the high incidence of tandem repeats in the virulence proteins of pathogens, toxins and allergens [3]. Genetic instability of these regions can allow a rapid response to environmental changes and thus can lead to emerging infection threats. In addition, the tandem repeats frequently occur in amyloidogenic, prion and other disease-related human sequences [4].

This presentation will provide a survey of current challenges in this area including identification of protein repeats in proteomes and structural prediction of the 3D structure of these proteins. The problem of identification of protein repeats is linked to the fact that, frequently, these repeats are strongly degenerated during evolution and, therefore, cannot be easily identified. To solve this problem, several computer programs which are based on different algorithms have been developed [1, 5, 6 7]. Nevertheless, there is still room for improvement of these methods.

Structural study of proteins with tandem repeats also represents a challenge. Most of these proteins have filamentous structure made of the repetition of equivalent modules [2]. Their large molecular weight and elongated filamentous shapes hamper conventional X-ray crystallography and NMR studies. As a result, these proteins are under-represented in the databases of the 3D structures. These difficulties increase the potential impact of hybrid approaches that combine bioinformatics analysis, molecular modeling, x-ray fiber diffraction, electron microscopy, STEM mass measurements, optical spectroscopy, and other biophysical techniques. In the past few years, a series of structural predictions have been made for such proteins including a beta-solenoid model of filamentous hemagglutinin (FHA) of *Bordetella pertussis*, alpha/beta-solenoid models of several Leucine-Rich-Repeat (LRR) proteins, an alpha-solenoid models of a human retromer protein VPS35 and subunits of proteasome, as well as superpleated beta-structures of prion and amyloid fibrils [8, 9, 10, 11, 12]. Some of the predicted molecular structures were subsequently determined experimentally confirming that the structures of proteins with tandem repeats can be predicted correctly. These examples suggest that, in general, *ab initio* prediction of such proteins is more reliable than prediction of globular proteins and this approach can be actively used for the structural annotations of genomes. Further development of reliable hybrid methods for prediction of the 3D structures of proteins with tandem repeats promise to be fertile research subjects of structural biology and bioinformatics.

References

[1] E.M. Marcotte, Pellegrini M, Yeates TO, Eisenberg D., A census of protein repeats. J. Mol. Biol., 293, 151–160, 1999.

[2] A.V. Kajava, Proteins with repeated sequence: structural prediction and modeling. J. Struct. Biology, 134:132-144, 2001.

[3] A.V. Kajava, J.M. Squire, D.A. Parry, Beta-structures in fibrous proteins.. Adv Protein Chem.73, 1-15, 2006.

[4] U. Baxa, T. Cassese, AV Kajava and AC Steven. Structure, function, and amyloidogenesis of fungal prions: filament polymorphism and prion variants. Adv Protein Chem. 73,125-180, 2006.

[5] A. Heger, and L. Holm, Rapid automatic detection and alignment of repeats in protein sequences. Proteins, 41, 224–237, 2000.

[6] A.M. Newman, and J.B. Cooper, XSTREAM: a practical algorithm foridentification and architecture modeling of tandem repeats in protein sequences. BMC Bioinformatics, 8, 382, 2007.

[7] J. Jorda and A.V. Kajava T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. Bioinformatics. 25, 2632-2638, 2009.

[8] A.V. Kajava, N. Cheng, R. Cleaver, M. Kessel, E. Willery , F. Jacob-Dubuisson, C. Locht and A.C. Steven, Beta-helix Model for the Filamentous Hemagglutinin Adhesin of *Bordetella pertussis* and Related Bacterial Secretory Proteins. Molecular Microbiology 42, 279-292, 2001.

[9] A.V. Kajava, What curves alpha-solenoids: Evidence for an alpha-helical toroid structure of Rpn1 and Rpn2 proteins of the 26S proteasome. J.Biol. Chem. 277: 49791-49798, 2002.

[10] A.V. Kajava and B. Kobe, Assessment of the ability to model proteins with leucine-rich repeats in light of the latest structural information. Protein Sci. 11: 1082-1090, 2002.

[11] A.V. Kajava, U. Baxa, R.B. Wickner and A. C. Steven, A Model for Ure2p Prion Filaments and Other Amyloids: The Parallel Super-pleated Beta-Structure Proc. Natl. Acad. Sci. USA, 101:7885-7890, 2004.

[12] A. Hierro, A.L. Rojas, R. Rojas, N. Murthy, G. Effantin, A.V. Kajava, A.C. Steven, J. S. Bonifacino, and J. H. Hurley, Functional architecture of the retromer cargo-recognition complex Nature 449:1063-1067, 2007.

Evaluating Genome Browsers Using a Software Qualification Method

Thomas Lacroix[1], Valentin Loux[1], Annie Gendrault[1], Jean-François Gibrat[1] and Hélène Chiapello[1]

[1] INRA UR1077, Unité Mathématique, Informatique & Génome, Jouy-en-Josas, France
{thomas.lacroix, valentin.loux, annie.gendrault, jean-francois.gibrat, helene.chiapello}@jouy.inra.fr

Abstract

Background: Because Genome Browsers (GBs) hold a central place in genomic projects, the diversity of tools available to scientists for visualizing and exploring genomes has increased dramatically over the last years. It often turns out to be a daunting task to compare and choose a well-adapted GB, as multidisciplinary knowledge is required to carry out this task and the number of tools, functionalities and features are overwhelming. To help the interested community making informed choices, there is an urgent need to apply and adapt standard software evaluation processes to bioinformatics tool families, such as GBs.

Results: We have implemented an industry promoted software qualification method, QSOS, to evaluate many of the available GBs using more than 120 criteria. We have defined about half of those criteria specifically for GBs, and incorporated the other half directly from QSOS's generic section. We have evaluated six GBs according to this methodology and present here a subset of our results organized according to three different user profiles: a biologist whose interest primarily lies into user-friendly and informative functionalities, a bioinformatician who wants facilities to integrate the GB into a wider framework, and a computer scientist who might choose a GB according to more technical features, for instance the possibility of developing a customized version by modifying the source code.

Conclusions: A website is publicly available at the URL http://genome.jouy.inra.fr/CompaGB. It offers a dedicated framework for GBs evaluation and comparison. It has been set up to help scientists to (1) choose GBs that would better fit their particular project, (2) visualize GBs features comparatively with easily accessible formats, such as tables or radar plots and (3) perform their own evaluation against the defined criteria.

Keywords : Genome Browser, software evaluation methodology, software comparison.

1 Background

Data in the field of genomics are ever growing in size and diversity, thus making visualization and data mining increasingly challenging. Advanced visualization tools such as genome browsers (GBs) have been developed to help biologists focusing on relevant clues with respect to their field of interest. The development of such complex software is often challenging and time consuming, but many genome browsers have been developed since the dawn of the 21st century: Bluejay [1], GenoMap [2], GenomeComp [3], GenomeViz [4] to cite a few. Projects aiming at supplanting or complementing current GBs are blooming as well. Although these different GBs provide the basic functionalities for browsing annotations on a genomic scaffold, their philosophy, functionality, interoperability and implementation are often unique or dedicated to a particular scientific field. The starting point for most of them lies in specific needs for a lab to study a particular problem and as a result, the current landscape of GBs is fragmented [5]. Such a disparity makes direct comparison difficult and external labs interested in integrating an existing GB to their own projects often ends up making their choice based on arbitrary decisions.

Attempts to categorize and compare GBs features have already been made, thus highlighting the need for guidance and clarity in this matter. In 2006, TS Furey carried out an overview and comparison of the UCSC Genome Browser, the Ensembl Genome Browser and the NCBI MapViewer along three axes: presentation, content and functionality [6]. In 2007, JD Gans and M Wolinsky published a summary table comparing 31 Genome Viewers according to 11 generic features, such as input formats or availability of the source code [7]. In this paper we implement a robust and traceable methodology to help creating a framework for GBs evaluations and comparisons. To the best of our knowledge, this is the first effort to compare GBs features using an open source standard

methodology and to set up a community resource centered on this kind of information. Here we present some results concerning the methodology, the criteria and the evaluation of six GBs according to three different user profiles: a biologist, a bioinformatician and a computer scientist. The six tools were chosen to cover a broad variety of tools, from the simplest locally developed GB to the most sophisticated one, disseminated into a large community of scientific labs and users around the world. We also included one recent tool from the last generation of GBs which was developed with Ajax technologies. Results are discussed in the light of those three distinct user contexts.

2 Methods

2.1 The QSOS methodology

The Qualification and Selection of Open Source software (QSOS) method [10] is designed to qualify, select and compare free and open source software in an objective, traceable and argued way. QSOS provides tools for defining the list of criteria, evaluating software, and a web server to visualize and compare the evaluations as a table or radar graph. It offers the possibility to weight the criteria so that they fit user specific context. For example, a criterion can be given a weight of 0 to indicate it is "unimportant", a weight of 1 if it is considered of average importance and a weight of 3 if it is critical for the user. The scoring is modulated according to the user's weighting to propose a selection of solutions that best meet the user requirements. The QSOS methodology version 1.6 provides a mandatory generic section which includes 65 criteria whose objective is to evaluate the potential and ease of integration of the software into a project: Who developed the software? What type of license for distribution? What is the richness of support / training / documentation? What is the frequency for releases?

2.2 Choice of six compared and evaluated Genome Browser

We have chosen to evaluate six GBs with this methodology: MuGeN (version 20060919) [8], GBrowse (version 1.69) [11], UCSC genome browser (version hg19) [12], Ensembl (version r54) [13], Artemis (version 10.08) / ACT (version 7.5.2) [14] and JBrowse [15]. GBs were chosen such as to cover a broad variety of software and functionalities: from a simple and easily accessible software developed by a local INRA team (MuGeN) to a representative selection of the most popular GBs that are potentially of interest for us. This selection is by no mean exhaustive and does not reflect the complete richness of the visualization tools in the field. It is intended to be a representative sample of the different types of GBs.

3 Results

3.1 The list of criteria

Comparing software packages with the QSOS methodology, which will be detailed later on, relies in part on scoring a set of weighted criteria. To provide the versatility required by the users with respect to the features of a GB, we built a list of criteria as comprehensive as possible. Our first step was to formulate a list of about 60 criteria tailored for GBs specificities. Three levels of scoring (full, limited/medium and poor) were defined specifically for each criterion to discriminate as objectively as possible between the different capabilities (see http://genome.jouy.inra.fr/CompaGB). Some criteria are for information purpose only (no score), e.g. the type of application. The criteria have been organized in four sections:

- **Section 1: Generic features**, which includes criteria concerning the origin, functionalities and the context of development of the GBs;

- **Section 2: Technical features**, which includes criteria such as the type of application, ease of installation, performance, supported platforms, API / interoperability, security...

- **Section 3: Data content and connectivity**, which focuses on criteria such as the possibility to display private data on top of public annotation, supported formats, connectivity with databases or web services, export features, data mining, ...

- **Section 4: Graphical User Interface (GUI),** that deals with criteria such as visualization techniques, richness of widgets, degree of customization for the tracks representing the genomic annotations, ease of navigation, comparative genomic features.

Finally, a more specific section entitled "**Annotation editing and creation**" is meant to be informative on the possibility of collaborative annotation, function assignment using a controlled ontology and assessment of the quality of the annotation. It was attached arbitrarily as a subsection of Section 3 (see Discussion).

Table 1 presents some criteria of section 1 - generic features - for the six evaluated Genome Browsers of this work: MuGeN [8], GBrowse [11], UCSC genome browser [12], Ensembl [13], Artemis/ACT [14] and JBrowse [15].

	Copy--right owner	Date publi--cation	Type appli--cation	Techno--logy	Main scientific purposes
MuGeN	INRA	2003	Stand alone	Perl - Gtk	Explora--tion B&LE*
Ensembl	EBI, EMBL, WTSI	2002	Web app	Perl CGI	Explora--tion, H&HE*
Artemis / ACT	WTSI	2000	Stand alone	Java	Creating annota--tion B&LE*
GBrowse	CSHL (NY), UC Berkley	2002	Web app	Perl CGI	Explora--tion MO*
UCSC	UC Santa Cruz	2002	Web app	C	Explora--tion, H&HE*
JBrowse	UC Berkley	2009	Web app	Ajax	Explora--tion MO*

Tab. 1. Example of generic features for the five compared GBs. * Abbreviations : B&LE : Bacteria and lower eukaryotes ; H&HE : Human and higher eukaryotes ; MO : Model organisms

3.2 The user profiles

We introduce three typical user profiles to illustrate our approach and present our evaluation for distinct needs and contexts.

3.2.1 The biologist profile

This first profile portrays a biologist, Dr Emma "doc" Brown, with no background in computer science and who would like to browse existing data on two model organisms, the human and the mouse. Dr Brown's field of research is Down syndrome and her lab aims to list genes whose increased expression is involved in perturbation of learning and memory faculties. She decides to use a GB to collect known information about the functions of chromosome 21 genes and the pathways in which they participate. The gene candidates will then be the focus of a series of molecular biology experiments using mouse models. In Table 2, we present some features that are on the wish list of users who, just like Dr Brown, want to browse and rapidly analyse both public and private data for a well defined purpose.

	Ergo--nomics	Search by sequence similarity	Search by annota--tion	Export graphic (vector / bitmap)	Simul--taneous views of multiple scales
MuGeN	+++	-	+	+++ / +++	-
Ensembl	+++	+++	+++	+++ / +++	+++
Artemis / ACT	+++	+++	+++	+ / +++	+
GBrowse	+++	+++	+++	+++ / +++	+++
UCSC	+++	+++	+++	+++ / -	+
JBrowse	+++	-	+++	- / -	-

Tab. 2. Example of criteria related to the ease of utilization

Though the ergonomics of all the interfaces among our selection are convenient, we conclude that Ensembl, JBrowse and Gbrowse provide the most user-friendly experience in terms of user support, responsiveness of the application, search functionalities and export facilities. The ability to export quality pictures in either variable or fixed resolution is a valuable asset to share findings in a publication. Ensembl make use of a fair number of technologies to speed up performances (mod_perl, Memcached and Ajax). JBrowse features Ajax technologies and was build from the ground up to offer a smooth and animated panning, zooming, navigation, and track selection. Its ergonomics and responsiveness are very user friendly but it is trailing GBrowse and Ensembl in term of functionalities.

Beyond those generic features, the use of a GB for a research biologist can vary greatly depending on the project goals and backgrounds. Use cases may include, but are not limited to, viewing heterogeneous information on gene annotation (ESTs, microarray data, comparative genomics,...), searching for known or disease-related genes, looking for variations in the genome (SNPs), browsing for sequence similarity with other species, generating illustrations for a publication,…. Some criteria were specifically formulated for those

different uses. Artemis is the only tool in our selection that offers advanced annotation and editing features.

3.2.2 The bioinformatician profile

The second scenario illustrates the needs of Dr Frank Hunstein, a bioinformatician that intends to use a GB into the framework of a wider pipeline and needs a local installation of the GB to support a large amount of private data sets from experimental labs. For example, a wet lab has successfully used tiling-array to study gene expression for the entire genome of their favorite organism. They wonder whether these exciting data can shed light on previously unidentified genes and Dr Hunstein offers to help them. Dr Frank Hunstein is used to writing scripts to format the data appropriately and join scattered bits of data but doesn't want to commit himself to decipher the source code. In his search for a GB, he is interested in looking at technical features relevant to connecting the different heterogeneous data sources and functionalities into a coherent construction. Table 3 presents some of these features for the six compared GBs.

	Upload private data	Connectivity to local files	Connectivity to databases
MuGeN	+	+	+
Ensembl	+++	+	+
Artemis / ACT	+	+	+
GBrowse	+++	+	+++
UCSC	+++	+++	+
JBrowse	+	+++	+

Tab. 3. Example of criteria related to the connectivity to a pipeline of data

We think DAS (Distributed Annotation System) protocol is a concept of interest for Dr Hunstein. It is a communication protocol used to exchange annotations on genomic or protein sequences and its philosophy is that annotations should not be provided by single centralized databases but spread over different locations. The GB then gathers up information from multiple Internet resources and integrates them into a single display to the user. Ensembl and Gbrowse support DAS protocol. Of importance to Dr Hunstein as well, are some criteria dealing with the process of installing the software and their local dependencies. Depending on the amount of annotation to support, the disk space required by the GB can quickly become trivial compared to the disk space needed by the annotation data.

3.2.3 The computer scientist profile

This third scenario presents a developer, Dr Rocky Auror, who wishes to have a customized version of an existing GB. Dr Auror is part of the "functional and evolutionary genomics" division of a DNA sequencing center, which provides the scientific community with access to high-throughput sequencing. Metagenomics projects producing a profile of bacterial diversity from soil samples have cluttered up his disk space with thousands of sequenced genomes. A bioinformatics tool that allows exploration of a vast amount of environmental samples is needed in order to carry out data mining and Dr Rocky Auror pictures a revolutionary navigator to show and explore thousands of genomes from a functional and evolutionary perspective. He decides to check out existing tools to serve as a solid basis for his creation. Some of our criteria are specifically formulated toward this kind of information, as table 4 shows.

	Modularity of the source code	Licensing's permissiveness	Adaptibility of the GUI to represent the annotation
MuGeN	+++	- (GPL)	+++
Ensembl	+++	+++ (Apache)	+++
Artemis / ACT	+++	- (GPL)	+++
GBrowse	+++	+++ (Artistic)	+++
UCSC	+++	+ (free for academic)	+++
JBrowse	+++	+++ (GNU LGPL and artistic)	+++

Tab. 4. Example of criteria related to software adaptability

Dr Rocky Auror also would like to further manipulate the data, run some analysis scripts or use an API for the purpose of data mining. Among the GBs that we have evaluated, Ensembl is the only one providing a Perl API that serves as a middle-layer between an application and the underlying database. With regards to data mining tools, both UCSC ("Table Browser") and Ensembl ("BioMart") have advanced user-friendly interfaces. Ensembl, Gbrowse and MuGeN have been developed with technical adaptability in mind, such as extending the source code by writing plug-ins. Gbrowse has a complete collection of analysis plug-ins, written by many different members of the GMOD community, such as PrimerDesigner, RestrictionAnnotator, Spectrogram, CreateBlast DB,... JBrowse is the most recent GB: it provides excellent performances in term of load time and memory use but it doesn't yet include such a diverse collection of plug-ins. Finally, Gbrowse, JBrowse and Ensembl offer the user the possibility to represent the data as a color gradient where the color intensity vary according to quantitative data.

3.3 The six compared GBs

Figures 1a and 1b summarize the scoring of our evaluations (without weighting of the criteria) for the six compared GBs.

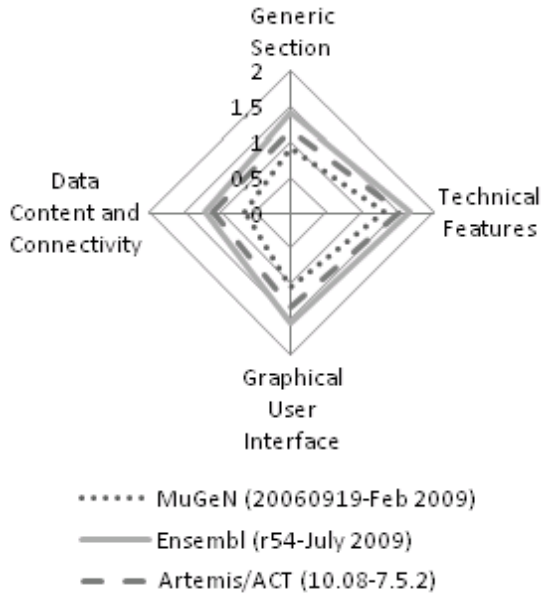


Fig. 1. Fig 1a. Radar plot averaging the four main section of criteria for MuGeN, Ensembl and Artemis/ACT

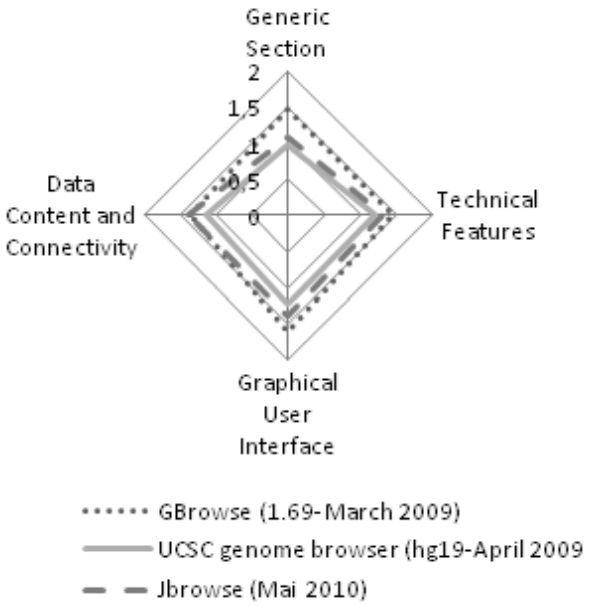


Fig 1b. Radar plot averaging the four main section of criteria for GBrowse, UCSC and JBrowse

Overall we found that MuGeN is a great tool to read and display a GenBank file, though its navigation features could be more complete. It is time and memory consuming for MuGeN to load higher eukaryote genomes. The UCSC browser natively displays a broad range of annotation on human, including cross-species comparisons but to the best of our knowledge, the underlying strategy lies into centralizing data on their server and no external labs have installed it locally for the purpose of storing and browsing their own data. The Ensembl website also provides exhaustive annotation for human and higher eukaryotes but also promotes itself has a flexible and open source software/database system that can be customized and locally installed. In this latter category, Gbrowse is very popular as well and has been designed to be an open source portable toolkit. Its strength lies in its anchoring into bioinformatics standards (DAS protocol, GFF format,…) and it can be easily customized to fit the needs of a model organism's community. JBrowse is an Ajax web application that features a very fast and ergonomic user interface. Even though it supports the same data sources (GFF, BED, WIG, SAM/BAM...) as GBrowse, JBrowse's development is more recent therefore its functionalities are yet less abundant. Artemis is the only choice among our selection for editing and creating a new set of annotations. With regards to GBs implementation strategies, we found that web applications offer the best balance between

deployment, accessibility, sharing of the data, and ease of connection to other web resources. Gbrowse or Ensembl offer the possibility to be locally installed, thus ensuring privacy, and/or to be installed as a public web-site.

4 Discussion and conclusion

4.1 Evaluation of GBs

This work was performed to provide clues to the community for performing well-argued and traceable GB evaluations. This does not mean that the used methodology makes GB evaluations totally unbiased. QSOS does not allow concurrent evaluation of the same software, and we think this might introduce a bias. Once we did perform a draft of an evaluation, we sent it to the team that developed the software for correction and comments. But even with this safeguard, evaluations reflect to some extent the perception and the background of the evaluators. To overcome this limit and ponder our evaluation so as to reflect a large consensus, we believe the only way is to have multiple concurrent evaluations from a larger community. In this regard we have set up a web site with forum, wiki, trackers, news, guidelines… where we allow users to post comments about previous evaluations, register to carry out an evaluation or suggest modifications to the list of criteria. As the project is open and anyone can participate and contribute evaluations, we also wrote a guideline that documents the whole process.

It is important to point out that the aim of this work is to improve the quality, the richness and reliability of evaluations. One consequence of this choice is that the evaluation process might appear quite time-consuming to some participants. Though it is very dependant of the evaluator background and the GB complexity, this might be dissuasive. This also explains the limited number of GBs evaluations our group has performed until now. We are aware of this weakness and we are planning to propose a simplified evaluation process in the next future.

4.2 The web site

We designed the web site http://genome.jouy.inra.fr/CompaGB as a framework dedicated to present and future GBs evaluations and comparisons. This resource can serve as a basis to get relevant information about GBs features and to make their comparisons easier. We encourage users of GBs to evaluate their software using this framework. We believe all public genome browsers have their strengths and weaknesses depending on what context and purpose it is to be used.

4.3 The QSOS methodology

The QSOS methodology helped us to frame and standardize GB comparisons. A weak point of the methodology is the definition of a list of relevant criteria, a step that can be very time consuming. Moreover, some specific criteria, such as “annotation editing and creation” do not fit easily into the main generalist sections. Finally, we decided to put the “annotation editing and creation” under the “Data content and connectivity” section of criteria but this in an arbitrary choice. We found also preferable to balance the depth of the list so that it is comprehensive yet not too tedious in order to keep the evaluation process as lightweight as possible. Concerning QSOS tools, one drawback we found to the web application is that it doesn't display scoreless criteria.

4.4 Challenges

A recent review assessed the strengths and limitations of the current genomic data visualization tools, and also the coming challenges in this dynamic field [9]. The authors pointed out four directions for software developers in order to meet future needs in genomics: enhancing data integration facilities, handling visual representation of comparisons of huge amounts of data (typically millions of elements), developing interactive interfaces allowing seamless navigation across relevant levels of resolution, and improving integration between automated computation and visualization. We wholly agree with these future needs, except that we add a supplementary one: the development of standard procedures for software evaluation and comparison, in order to help the scientific community to perform consistent choices.

Authors’ contribution

TL carried out the study, performed GBs evaluations and comparisons and drafted the manuscript. VL, HC, JFG and AG participated in the design of the QSOS criteria and in the results analysis and interpretation. HC initiated the study, performed some GBs evaluations and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing help and support. We thank P. Bessières, C. Caron, M. Hoebeke and S. Sidibé-Bocs for helpful discussions.

References

[1] Soh J, Gordon PM, Taschuk ML, Dong A, Ah-Seng AC, Turinsky AL, Sensen CW: Bluejay 1.0: genome browsing and comparison with rich customization provision and dynamic resource linking. BMC Bioinformatics 2008, 9:450.

[2] Sato N, Ehira S: GenoMap, a circular genome data viewer. Bioinformatics 2003, 19(12):1583-1584.

[3] Yang J, Wang J, Yao ZJ, Jin Q, Shen Y, Chen R: GenomeComp: a visualization tool for microbial genome comparison. J Microbiol Methods 2003, 54(3):423-426.

[4] Ghai R, Hain T, Chakraborty T: GenomeViz: visualizing microbial genomes. BMC Bioinformatics 2004, 5:198.

[5] Stein L: Creating a bioinformatics nation. Nature 2002, 417(6885):119-120.

[6] Furey TS: Comparison of human (and other) genome browsers. Hum Genomics 2006, 2(4):266-270.

[7] Gans JD, Wolinsky M: Genomorama: genome visualization and analysis. BMC Bioinformatics 2007, 8:204.

[8] Hoebeke M, Nicolas P, Bessieres P: MuGeN: simultaneous exploration of multiple genomes and computer analysis results. Bioinformatics 2003, 19(7):859-864.

[9] Nielsen CB, Cantor M., Dubchak I., Gordon D., T. W: Visualizing genomes: techniques and challenges. Nature Methods Supplement 2010, 7(3s):S1-S11.

[10] Method for Qualification and Selection of Open Source software (QSOS) [www.qsos.org]

[11] Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A et al: The generic genome browser: a building block for a model organism system database. Genome Res 2002, 12(10):1599-1610.

[12] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: The human genome browser at UCSC. Genome Res 2002, 12(6):996-1006.

[13] Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L et al: Ensembl 2009. Nucleic Acids Res 2009, 37(Database issue):D690-697.

[14] Carver T, Berriman M, Tivey A, Patel C, Bohme U, Barrell BG, Parkhill J, Rajandream MA: Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics 2008, 24(23):2672-2676.

[15] Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: A next-generation genome browser. Genome Res. (2009).

Web Services for MIcrobial Genome Annotation using Data Integration

Célia MICHOTEY[1], Ludovic LEGRAND[1], Hélène CHIAPELLO[2], Valentin LOUX[2], Annie GENDRAULT[2], Jean-François GIBRAT[2] and Christophe CARON[1]

[1]Station Biologique, FR2424, Plateforme ABIMS, CNRS-UPMC, Place G.. Teissier, 29680 Roscoff Cedex, France
{celia.michotey, ludovic.legrand, christophe.caron}@sb-roscoff.fr

[2]Unité Mathématique, Informatique & Génome, UR 1077 INRA, Domaine de Vilvert, 78352, Jouy-en-Josas, France
{helene.chiapello, valentin.loux, annie.gendrault, jean-francois.gibrat}@jouy.inra.fr

Keywords: annotation, database, integration, Web Service, comparative genomics, proteomics.

Annotation de génomes microbiens via l'intégration de données

.

Mots-clés annotation, bases de données, intégration, Service Web, génomique comparée, protéomique.

1 Introduction

L'annotation structurale et fonctionnelle des génomes constitue souvent la première étape nécessaire à la description et à la compréhension des fonctions biologiques présentes chez un organisme microbien. Cependant, le processus standard d'annotation d'une souche bactérienne isolée n'offre souvent que des réponses partielles quant aux spécificités fonctionnelles et physiologiques de l'organisme concerné. Plusieurs publications récentes montrent l'intérêt d'une stratégie dite « protéogénomique » (intégration massive de données protéomiques et génomiques) pour comprendre les spécificités fonctionnelles (adaptation, pathogénicité) de certaines espèces ou souches bactériennes [1, 2].

Pour élaborer des stratégies d'annotation innovantes basées sur l'intégration de sources d'informations complémentaires mais hétérogènes, deux grands types de méthodes existent : l'intégration physique des données (avec des applications type entrepôt) et l'intégration virtuelle, basée sur des protocoles de communication entre applications permettant un accès transparent.

Dans ce travail, nous proposons une méthodologie basée sur l'intégration virtuelle d'un environnement logiciel existant pour l'annotation des génomes microbiens, AGMIAL [3], avec des outils plus ciblés sur deux domaines qui nous paraissent majeurs pour la génomique microbienne : la génomique comparée, via l'application MOSAIC [4] et la protéomique expérimentale, via l'application PARIS [5].

2 Méthodologie

Du point de vue technique, notre choix s'est porté sur les Services Web qui ont pour avantage de pouvoir s'implémenter a posteriori autour des applications. Ce sont des technologies bien adaptées au déploiement et à l'intégration de composants logiciels dans un environnement hétérogène, et qui permettent aussi de gérer la diversité des architectures.

Les trois composants logiciels ne disposant pas de la même architecture logicielle, différentes étapes ont été nécessaires pour faciliter leur intégration :

1. Nous avons tout d'abord associé à chaque composant des scénarios liés au processus.
2. Il a ensuite s'agit de rendre chaque composant fournisseur de données en développant une couche de Services Web avec une API (Application Program Interface) homogène.
3. Enfin nous avons développé des interfaces de représentation des données dans chaque application afin de rendre les applications potentiellement clientes des autres composants.

3 Résultats

Une dizaine de scénarios biologiques pour l'aide à l'annotation ont été construits, et trois ont été

considérés par les biologistes comme prioritaires (voir table 1 ci dessous).

(Scénario) Question biologique	AGMIAL	MOSAIC	PARIS2	Contraintes
(1) Aide à l'annotation automatique d'un génome dans AGMIAL (régions conservées/variables d'une souche par rapport à une souche de référence)	CO	FO	-	- Génome de référence existant - Gestion de la confidentialité
(2) Aide à l'interprétation des protéomes expérimentaux dans PARIS (gels de 2 souches de la même espèce)	CO	FO	CO	- Gestion de la confidentialité
(3) Aide à l'annotation de protéines d'intérêt dans AGMIAL (spot correspondant présent dans un gel protéique de PARIS)	CO	-	FO.	- Gestion de la confidentialité

Tab. 1. Exemples de trois scénarios développés pour construire l'intégration des 3 applications AGMIAL, MOSAIC et PARIS. (CO=Consommateur, FO=Fournisseur)

Ces trois scénarios sont en cours d'implémentation au niveau des interfaces des applications AGMIAL et MOSAIC. Ils vont être testés sur des données biologiques de *Propionibacterium freudenreichii* et *Lactococcus lactis*. Ces Services Web apportent une aide précieuse pour la validation des annotations de l'application AGMIAL, qui reste le composant intégrateur de la nouvelle architecture. Cependant, nos scénarios montrent que les autres applications peuvent également bénéficier de cette interopérabilité.

Par exemple, la classification des protéines d'un génome donné dans AGMIAL ou PARIS peut se faire en fonction du type de segment, variable (spécifique à une souche donnée) ou conservé, prédit par MOSAIC, sur lequel se situent leurs gènes.

Au final les Services Web développés, et basés sur les protocoles SOAP et HTTP, se révèlent être une technique peu intrusive et donc assez légère à mettre en place. Ce choix nous a permis, de plus, d'intégrer des Services Web externes au projet, comme ceux de l'application Kegg (http://www.genome.jp/kegg/soap/) pour les voies métaboliques. Cependant, nous avons été confrontés à différents problèmes, en particulier, la gestion des données privées qui nécessite une réflexion complémentaire autour de la sécurité et notamment l'homogénéité des processus d'authentification à travers les applications.

4. Conclusion

Une des limites importantes de l'intégration de données concerne souvent les problèmes sémantiques qui nécessitent un travail de modélisation et de mise au point d'ontologies. Pour s'en affranchir partiellement, l'intégration virtuelle d'objets biologiques simples et bien identifiés que nous avons réalisée autour des trois applications AGMIAL, MOSAIC et PARIS constitue un moyen rapide, efficace et peu intrusif de faire des liens entre les analyses *in silico* et des données expérimentales, tout en enrichissant le processus d'annotation. Ce type d'architecture distribuée est robuste vis-à-vis de l'évolution spécifique de chaque application. Cette première intégration lève donc un verrou technologique et va permettre d'élaborer des scénarios incluant des relations plus complexes entre les trois composants et des ressources externes. Enfin, nous souhaitons enrichir ce système avec l'intégration d'autres sources données, et plus particulièrement les données de transcriptomique.

Remerciements

Ce travail est financé par le projet ANR PFTV 2007 (projet MIGADI 2008-2010). Les auteurs remercient l'UMR1253 de l'INRA de Rennes (H. Falentin, G. Jan, S-M. Deutsch).

Références

[1] de Groot A, *et al*. Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium Deinococcus deserti. PLoS Genet 2009, 5(3):e1000434.

[2] Lamontagne *et al*. Proteomics-based confirmation of protein expression and correction of annotation errors in the Brucella abortus genome. BMC Genomics 2010, 11:300.

[3] K. Bryson, V. Loux, R. Bossy, P. Nicolas, S. Chaillou, M. van de Guchte, S. Penaud, E. Maguin and JF. Gibrat. AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Research. 2006, 34, 3533-3545.

[4] H. Chiapello, A. Gendrault, C. Caron, J. Blum, MA Petit and M El Karoui. MOSAIC: an online database dedicated to the comparative genomics of bacterial strains at the intra-species level. BMC Bioinformatics 2008, 9:498

[5] J. Wang, C. Caron, M.-Y. Mistou, C. Gitton and A. Trubuil. PARIS: a proteomic analysis and resources indexation system. Bioinformatics 2004, 20:133-135.

Exact distribution of a pattern in a set of random sequences

G. NUEL[1], L. REGAD[2], J. MARTIN[3] and A.-C. CAMPROUX[2]

[1] MAP5, Department of Applied Mathematics, CNRS UMR-8145, Paris Descartes University, France
gregory.nuel@parisdescartes.fr
[2] MTi, *Molécules Thérapeutiques in silico*, INSERM UMRS-973, University Paris Diderot University, France
{leslie.regad@univ-paris-diderot.fr,
anne-claude.camproux@univ-paris-diderot.fr}@email.fr
[3] IBCP, *Institut de Biologie et Chimie des Protéines*, IFR 128, CNRS UMR 5086, University of Lyon 1, France
juliette.martin@ibcp.fr

Keywords Markov chain embedding (MCE), Deterministic Finite Automaton (DFA), heterogeneous Markov models, PROSITE signature, regulation motifs

1 Introduction

In bioinformatics it is common to search for a pattern of interest in a potentially large set of rather short sequences (upstream gene regions, proteins, exons, etc.). Although many methodological approaches allow practitioners to compute the distribution of a pattern count in a random sequence generated by a Markov source (see [1,2,3] for a review), no specific developments have taken into account the counting of occurrences in a set of independent sequences. We aim to address this problem by deriving efficient approaches and algorithms to perform these computations both for low and high complexity patterns in the framework of homogeneous or heterogeneous Markov models. This work has been published in AMB.

2 Methods

Model and notations Let $(X_i)_{1\leqslant i\leqslant \ell}$ be an order $d \geqslant 0$ Markov chain over the finite alphabet $\mathcal{A}$ (with cardinal $|\mathcal{A}| \geqslant 2$). For all $1 \leqslant i \leqslant j \leqslant \ell$, we denote by $X_i^j = X_i \dots X_j$ the subsequence between positions i and j. For all $a_1^d = a_1 \dots a_d \in \mathcal{A}^d$, $b \in \mathcal{A}$, and $1 \leqslant i \leqslant \ell - d$, let us denote by $\mu\left(a_1^d\right) = \mathbb{P}\left(X_1^d = a_1^d\right)$ the starting distribution and by $\pi_{i+d}(a_1^d, b) = \mathbb{P}(X_{i+d} = b | X_i^{i+d-1} = a_1^d)$ the transition probability towards X_{i+d}.

Let $\mathcal{W}$ be a finite set of words over $\mathcal{A}$ (for simplification purpose, we assume that $\mathcal{W}$ contains no word of length less than d – in the general case, one may have to count the pattern occurrences already seen in X_1^d, which results in a more complex starting distribution for our embedding Markov chain). We consider the random number N_ℓ of matching positions of $\mathcal{W}$ in X_1^ℓ defined by: $N_\ell = \sum_{i=1}^{\ell} \mathbb{I}_{\{\mathcal{W} \cap \mathcal{S}(X_1^i) \neq \emptyset\}}$ where $\mathcal{S}(X_1^i)$ is the set of all the suffixes of X_1^i and where $\mathbb{I}_A$ is the indicator function of event A.

DFA and Markov chain embedding As suggested in [4,5,6,7], we perform an optimal Markov chain embedding of our pattern problem through a Deterministic Finite Automaton (DFA). We use here the notations of [7]. Let $(\mathcal{A}, \mathcal{Q}, \sigma, \mathcal{F}, \delta)$ be a *minimal* DFA that recognizes the language $\mathcal{A}^*\mathcal{W}$ of all texts over $\mathcal{A}$ ending with an occurrence of $\mathcal{W}$. $\mathcal{Q}$ is a finite state space, $\sigma \in \mathcal{Q}$ is the starting state, $\mathcal{F} \subset \mathcal{Q}$ is the subset of final states and $\delta : \mathcal{Q} \times \mathcal{A} \to \mathcal{Q}$ is the transition function. Without loss of generality, we make the technical assumption that this automaton is non d-ambiguous (see [7] for details).

We then consider the random sequence defined over $\mathcal{Q}$ by $\widetilde{X}_0 = \sigma$ and $\widetilde{X}_i = \delta(\widetilde{X}_{i-1}, X_i)\ \forall i, 1 \leqslant i \leqslant \ell$. One can prove that this random sequence has two properties (see [7] for details):

i) $(\widetilde{X}_i)_{i \geqslant d}$ is a heterogeneous order 1 Markov chain over $\mathcal{Q}$ which starting distribution $\mathbf{m}_d$ and transition matrix $\mathbf{T}_{i+d}(p, q)$ can be expressed directly from μ, π_{d+i} and the DFA;
ii) $\mathcal{W} \cap \mathcal{S}(X_1^i) \neq \emptyset \iff \widetilde{X}_i \in \mathcal{F}$ which means that occurrences of $\mathcal{W}$ in X_1^ℓ can be tracked through occurrence of $\mathcal{F}$ in $\widetilde{X}_d^\ell$.

MGF for a set of sequences Let us now assume that we consider a set of r sequences. For any particular sequence j (with $1 \leqslant j \leqslant r$) we denote by ℓ_j its length, by N_{ℓ_j} its number of pattern occurrences, and by $\mathbf{m}_d^j$, $\mathbf{T}_{i+d}^j = \mathbf{P}_{i+d}^j + \mathbf{Q}_{i+d}^j$ (this decomposition store in $\mathbf{Q}_{i+d}^j$ only transitions towards a final state) its corresponding Markov chain embedding parameters.

If we denote by $G_N(y) = \sum_{n=0}^{+\infty} \mathbb{P}(N_{\ell_1} + \dots + N_{\ell_r} = n) y^n$ the Moment Generating Function (MGF) of $N = N_{\ell_1} + \dots + N_{\ell_r}$, we have: $G_N(y) = G_{N_{\ell_1}}(y) \times \dots \times G_{N_{\ell_r}}(y)$ with $G_{N_{\ell_j}}(y) = \mathbf{m}_d^j \left(\prod_{i=1}^{\ell_j - d} \left(\mathbf{P}_{i+d}^j + y \mathbf{Q}_{i+d}^j \right) \right) \mathbf{1}^{\mathrm{T}}$. One should note

that the formula is dramatically simplified in the homogeneous case where one gets $G_{N_{\ell_j}}(y) = \mathbf{m}_d^r (\mathbf{P} + y\mathbf{Q})^{\ell_r - d} \mathbf{1}^{\mathrm{T}}$. Efficient algorithms to compute these quantities are given in the AMB paper.

3 Applications

PROSITE signatures We now consider the complete proteome of the bacteria *Escherichia coli* (File NC_000913.faa). This data set encompasses a total of $4,131$ protein sequences with lengths ranging from 14 to $2,358$ aminoacids. We fit on this data set a homogeneous order 1 Markov model which is used to derive over-representation P-values of patterns (see Table Tab. 1).

PROSITE signature	L	n	exact
PILI_CHAPERONE	226	10	3.27×10^{-46}
SIGMA54_INTERACT_2	313	12	1.58×10^{-42}
EFACTOR_GTP	320	8	4.43×10^{-20}
ALDEHYDE_DEHYDR_CYS	331	11	5.63×10^{-9}
ADH_ZINC	478	12	8.93×10^{-16}
THIOLASE_1	637	5	5.76×10^{-9}
SUGAR_TRANSPORT_1	796	18	3.75×10^{-8}
FGGY_KINASES_2	2668	5	2.14×10^{-4}
PTS_EIIA_TYPE_2_HIS	2758	8	7.19×10^{-19}
MOLYBDOPTERIN_PROK_3	3907	11	2.59×10^{-35}
SUGAR_TRANSPORT_2	6689	10	1.22×10^{-5}

Tab. 1. Exact P-values for a selection of PROSITE patterns of high complexities. n is the number of observed occurrence, L is the DFA size.

Regulation motifs We retrieved the sequence of transcription factor binding sites of *Saccharomyces cerevisiae* on the YEASTRACT website and searched for a subset of these transcription factor binding sites in the upstream regions of yeast genes, retrieved on the Regulatory Sequence Analysis Tools website. This data set comprises a total of $1,371$ upstream sequences between positions -800 and -1 (the length is hence $\ell = 800$ for each sequence).

On these data, we fit an order 1 homogeneous Markov model as well as a heterogeneous Markov model of same order fitted using a classical sliding window approach (window size arbitrary set to 200). We then derive exact P-values (see Table Tab. 2).

4 Conclusion

The results presented here allow for the first time to compute exactly the distribution of a pattern in a set of random sequences by fully taking into account the fragmented structure of the problem. Thanks to efficient algorithms, it is possible both to deal with high complexity patterns in relatively large dataset (PROSITE signature) as well as to work with fully heterogeneous background models (regulation motifs). All these methods will be soon implemented in the SPatt package `http://stat.genopole.cnrs.fr/spatt`.

DNA pattern	n	L	homogeneous	heterogeneous
CGCACCC*	28	10	2.95×10^{-3}	3.74×10^{-3}
AAGAAAAA*	427	11	1.31×10^{-99}	1.29×10^{-99}
AACAACAAC	25	10	1.76×10^{-6}	1.38×10^{-6}
TCCGTGGA*	22	11	1.12×10^{-6}	1.55×10^{-6}
GCGCGCGC	18	11	6.52×10^{-10}	1.65×10^{-9}
RTAAAYAA*	391	14	7.70×10^{-12}	1.68×10^{-12}
WWWTTTGCTCR*	15	17	4.15×10^{-1}	4.09×10^{-1}
A{24}	42	27	2.05×10^{-23}	2.14×10^{-22}
TAWWWWTAGM*	212	36	3.08×10^{-9}	3.04×10^{-9}
YCCNYTNRRCCGN*	11	40	3.10×10^{-2}	3.05×10^{-2}
GCGCN{6}GCGC	1	106	8.97×10^{-1}	8.84×10^{-1}
CGGN{8}CGG*	102	183	1.26×10^{-14}	1.73×10^{-13}
GCGCN{10}GCGC	6	464	2.88×10^{-2}	2.84×10^{-2}

Tab. 2. Exact P-values for several DNA patterns (known transcription factors are marked with a star) in the upstream region data set using either a homogeneous or heterogeneous background model.

References

[1] M. Reignier, A unified approach to word occurrences probabilities. *Discrete Applied Mathematics*, 104(1):259-280, 2000.

[2] M. Lothaire, Applied Combinatorics on Words. *Cambridge University Press*, 2005.

[3] G. Nuel, Numerical solutions for Patterns Statistics on Markov chains. *Stat. App. in Genet. and Mol. Biol.*, 5(1):26.

[4] P. Nicodème, B. Salvy and P. Flajolet, Motif statistics. *Theoretical Com. Sci.*, 287(2):593–617, 2002.

[5] M. Crochemore and V. Stefanov, Waiting time and complexity for matching patterns with automata. INFO. PROC. LETTERS, 87(3):119–125, 2003.

[6] M. E. Lladser, Mininal Markov chain embeddings of pattern problems, *Information Theory and Applications Workshop*, 251-255, 2007.

[7] G. Nuel, Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. *J. of Applied Prob.*, 45(1):226-243, 2008.

Influence of the rearrangement rates on the organization of genome transcription

David P. PARSONS[1,3], Carole KNIBBE[2,3] and Guillaume BESLON[1,3]

[1] Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
[2] Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France
[3] IXXI, Institut Rhône-Alpin des Systèmes Complexes, Lyon, F-69007, France
{david.parsons, carole.knibbe, guillaume.beslon}@liris.cnrs.fr

Keywords Digital genetics, Operons, Chromosomal rearrangements, Transcription, Genome organization.

Influence des taux de réarrangements chromosomiques sur l'organisation de la transcription

Mots-clefs Génétique digitale, Opérons, Réarrangements chromosomiques, Transcription, Organisation des génomes.

1 Introduction

L'organisation des génomes présente une très grande diversité. D'un côté, la plupart des génomes viraux sont très courts et très denses, ne contenant que très peu de séquences non-codantes. À l'autre extrême, les génomes eucaryotes multicellulaires, très longs, contiennent une grande part de non-codant. Ces différences s'accompagnent de variations dans l'organisation de la transcription : les génomes les plus courts et les plus denses sont en général transcrits en de longs ARNs pouvant contenir plusieurs CDS (opérons) alors que les génomes longs donnent naissance à une pléthore d'ARNs qui contiennent rarement plus d'une CDS, la majorité n'en contenant aucun.

L'origine de ces différences est relativement mal connue. Parmis les hypothèses existantes pour expliquer cette diversité, nous nous intéressons ici à celle du fardeau mutationnel proposée par M. Lynch [1]. Selon cette hypothèse, l'ADN en excès est mutagène pour les séquences codantes voisines. Ainsi, lorsque le taux de mutations est fort, seuls les génomes compacts peuvent être transmis fidèlement. Cependant, l'étude expérimentale d'une telle hypothèse est difficile étant donnée la complexité des processus en œuvre et les échelles de temps sur lesquelles ils se déroulent. Les approches de génomiques comparatives permettent de contourner cette difficulté, cependant, elles sont basées sur un état figé des séquences et doivent *inférer* leur passé évolutif.

Les simulations *in silico* ont déjà montré leur fort potentiel dans ce type de questionnement, mettant en évidence des pressions indirecte qu'il aurait été difficile d'identifier autrement [2,3]. Elles permettent d'avoir une vue dynamique du processus évolutif en un temps raisonnable avec un contrôle fin des paramètres. Nous proposons ici d'explorer les effets des taux de mutations et de réarrangements sur l'organisation des transcrits en utilisant le modèle de génétique digitale Aevol [3,4].

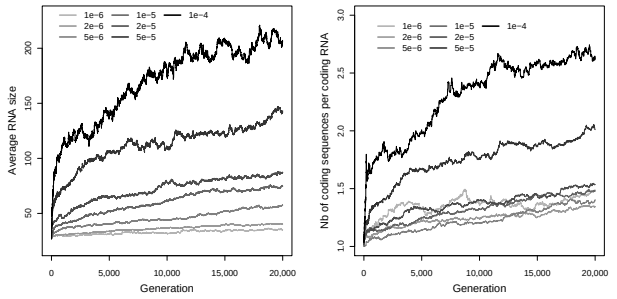
2 Résultats

Nous avons fait évoluer 147 populations de 1000 individus dans des environnement identiques et stables. Les seuls paramètres variant d'une simulation à l'autre sont les taux de mutations ponctuelles et de réarrangements, les valeurs testées s'étalant de 1.10^{-6} à 1.10^{-4} par bp (voir Fig. 1 et Fig. 2).

Les individus de la population, initialisée avec des génomes aléatoires, acquièrent progressivement de nouveaux gènes et les modifient de sorte que le protéome répond aux exigences de l'environnement. Après quelques milliers de générations, toutes les populations se sont adaptées à leur environnement, cependant, on observe que les stratégies évolutives qui ont émergé sont très différentes selon les simulations. On constate en particulier une grande diversité de tailles de génomes, celles-ci s'étalant de 1000 à 200.000 bp. Cette observation confirme les résultats obtenus dans [3], montrant une forte corrélation entre la taille des génomes et les taux de réarrangements chromosomiques, un fort taux de réarrangements entrainant une compaction du génome (et en particulier, des séquences non-codantes) du fait d'une pression indirecte pour un certain niveau de robustesse. Le

taux de mutations ponctuelles a une influence beaucoup plus faible sur la structure des génomes obtenus.

L'étude de la structure des transcrits montre que ces variations de taille s'accompagnent de profondes différences dans la façon dont les génomes sont transcrits. En effet, les génomes les plus longs présentent de très nombreux ARNs non-codants, leurs ARNs codants étant courts et ne contenant pour la plupart qu'une unique CDS. Les génomes courts, quant à eux, sont généralement transcrits en des ARNs beaucoup plus longs contenant pour la plupart plusieurs gènes, formant ainsi des opérons (Fig. 1).



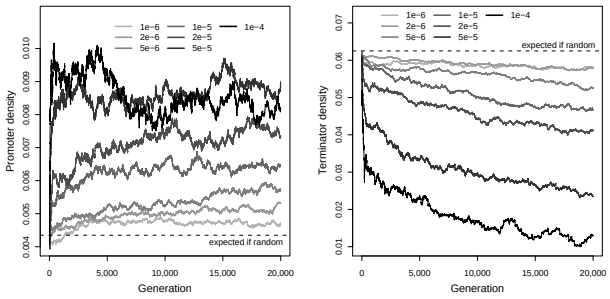
(a) Taille des ARNs (b) Nombre de gènes par ARN

Fig. 1. Évolution de la taille moyenne des ARNs (codants ou non-codants) et du nombre moyen de gènes par ARN codant (contenant au moins une CDS). Pour des raisons de lisibilité, les données présentées ici ont été agrégées, chaque ligne représentant la valeur moyenne des 21 simulations partageant le même taux de réarrangement.

La Fig. 1(b) montre que l'apparition d'opérons n'est présente qu'au delà d'un certain taux de réarrangements. Cet effet de seuil s'explique par l'effet combiné de deux pressions antagonistes. Selon l'hypothèse du fardeau mutationnel, seuls les génomes courts peuvent être transmis fidèlement lorsque le niveau de variations génétiques est élevé. Par ailleurs, la sélection des individus les plus adaptés à l'environnement tend ici à favoriser ceux ayant beaucoup de gènes. La conjonction de ces deux pressions résulte en une nouvelle pression sur la densité des génomes.

Lorsque le taux de réarrangements est modéré, la densité optimale peut être obtenue simplement en réduisant la taille du non-codant. Cependant, au delà d'un certain seuil, la réduction du non-codant devient insuffisante et il devient nécessaire de trouver d'autres moyens de compacter le génome. Dans nos simulations, cela se traduit par l'augmentation de la taille des transcrits, permettant à plusieurs gènes d'être présents sur le même brin d'ARN.

La Fig. 2 permet d'analyser la dynamique qui mène à cet allongement des ARNs. Seuls les terminateurs semblent être éliminés régulièrement tout au long de l'évolution, la densité de promoteurs demeurant stable. Or, les terminateurs morcellent le génome, créant des zones qui ne peuvent pas être traduites. En se débarrassant d'une partie de ses terminateurs, un individu peut donc optimiser son génome, le rendant plus compact tout en conservant une quantité de séquences codantes raisonnable.



(a) Densité de promoteurs (b) Densité de terminateurs

Fig. 2. Évolution de la densité moyenne de promoteurs (a) et de terminateurs (b) pour les différents taux de réarrangements.

Nos simulations reproduisent donc fidèlement les différences d'organisation des transcrits observées sur des organismes réels. La forte dépendance de la structure des transcrits aux taux de réarrangements est un argument fort en faveur de l'hypothèse du fardeau mutationnel, les individus soumis à un fort taux de réarrangements présentant un génome "optimisé" aussi bien sur le plan de la taille que sur l'organisation de la transcription. En outre, le modèle permet d'explorer la dynamique de cette optimisation. Ici, c'est le nombre de séquences terminatrices qui permet cette optimisation, ce qui engage à rechercher les traces de ce mécanisme dans les génomes réels. Enfin, cette première expérience nous permet d'envisager de nouveaux développements, par exemple en modifiant la taille de la population, qui est un autre paramètre important dans l'hypothèse du fardeau mutationnel [1].

Références

[1] M. Lynch, Streamlining and Simplification of Microbial Genome Architecture. *Annu. Rev. Microbiol.*, 60(1) :327-349, 2006.

[2] C. Adami, Digital genetics : unravelling the genetic basis of evolution. *Nat. Rev. Genet.*, 7(2) :109-118, 2006.

[3] C. Knibbe, A. Coulon, O. Mazet, J.-M. Fayard and G. Beslon, A long-term evolutionary pressure on the amount of noncoding DNA. *Mol. Biol. Evol.*, 24(10) :2344-2353, 2007.

[4] C. Knibbe, *Structuration des génomes par sélection indirecte de la variabilité mutationnelle, une approche de modélisation et de simulation*. PhD thesis, INSA-Lyon, 2006.

METEOR - a plateform for quantitative metagenomic profiling of complex ecosystems

Nicolas PONS [1], Jean-Michel BATTO, Sean KENNEDY, Mathieu ALMEIDA, Fouad BOUMEZBEUR, Bouziane MOUMEN, Pierre LEONARD, Emmanuelle LE CHATELIER, S. Dusko EHRLICH and Pierre RENAULT

[1] INSTITUT MICALIS, UMR1319 INRA, Domaine de Vilvert, 78352, Jouy-en-Josas, Cedex, France
nicolas.pons@jouy.inra.fr

The study of complex microbial ecosystems by a quantitative metagenomic approach has been made possible by advancements in high-throughput sequencing technologies. Quantitative metagenomics relies on deep sequencing to construct an ecosystem profile using gene and genome counts. Next generation sequencing (NGS) technologies such as SOLiD or Illumina produce millions of short sequences (35 to 75bp) which can be used as tags to establish gene profiles. This approach requires the use of a specific reference catalog which should be composed of genes present in the ecosystem of interest. The use of classical bioinformatic methods for the analysis of such large amounts of data is not feasible as we overpass the expected dataset size of common tools. We have therefore developed an integrated metagenomic analysis pipeline, METEOR, which includes the indexing of short reads to genomic objects. Data are indexed in an embedded database around the iMOMi framework [1] and organized in a dedicated file system. This optimization facilitates secondary analysis including gene/species abundance evaluation, cross-sample comparison, ecosystem metabolism reconstruction or gene/species diversity analysis. The METEOR pipeline has been implemented for several metagenomic projects such as MicroObes for characterizing the human intestinal microbiome of obese individuals following a restrictive diet or FoodMicrobiomes for studying the ecosystem of fermented food like French traditional cheeses.

In MicroObes, we investigate the changes of gut microbiota in a human model of weight loss induced by restrictive diet in moderately obese subjects. DNA isolated from 195 faecal samples of 49 obese subjects collected at different date (start of the study, 6 weeks after a restrictive diet and 12 weeks) have been sequenced with SOLiD technology yielding about 300 gigabases. Short reads have been indexed against the 3.3 millions genes of the human gut microbial gene catalog (MetaHIT consortium [2]). Statistical analysis of the gene profiles generated indicate significant variations in gene and genome frequencies during the first 6 weeks of dieting and a subsequent stabilization after 12 weeks according to the observed success of patients dietary.

References

[1] N. Pons, J.M. Batto, SD. Ehrlich and P. Renault, Development of software facilities to characterize regulatory binding motifs and application to streptococcaceae. J Mol Microbiol Biotechnol, 14(1-3): 67-73, 2008.

[2] J. Qin, ..., N. Pons, ..., J.M. Batto, ..., P. Renault, ..., S.D. Ehrlich, J. Wang. A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464(7285):59-65, 2010.

Prediction of patterns of interest from protein primary sequence through structural alphabet

Christelle REYNÈS[1], Leslie REGAD[1], Robert SABATIER[2] and Anne-Claude CAMPROUX[1]

[1] Unité MTi, U973 Inserm - Paris 7 Diderot, 35, rue Hélène Brion, 75205 Paris, Cedex 13, France
{christelle.reynes, leslie.regad, anne-claude.camproux}@univ-paris-diderot.fr
[2] Laboratoire de Physique Industrielle et Traitement de l'Information, EA 2415, Faculté de Pharmacie 15, avenue Charles Flahault, BP 14491, 34093 Montpellier Cedex 5, France
sabatier@univ-montp1.fr

Abstract *The prediction of patterns in proteins that have been identified as interesting (funtional or turns for example) is really important in the present context of high-throughput sequencing programs. The proposed method allows to predict such patterns once they have been identified as structural letter words through a Hidden-Markov model structural alphabet. Once a pattern is chosen, it can be predicted directly from sequence (depending on a certain sequence dependency) without knowing anything about 3D structure. It consists in two steps: firstly the dependencies between structural letters and amino-acids is learnt by Genetic Programming as boolean trees, secondly, information from the first step as well as from the dependencies between consecutive structural letters is combined by a Hidden-Markov model which allows to score the probability to find the target pattern given any amino-acid sequence. The method is illustrated on three different patterns related to ATP-, SAH/SAM-binding sites and to specific turns.*

Keywords sequence-based automatic annotation, hidden Markov models, structural alphabet, functional patterns.

1 Introduction

Mining data about protein sequences is of prime importance as sequencing technologies are constantly providing new amino-acid (AA) sequences with often few functional knowledge. For example, UniProtKB (release 2010_04, Mar 19, 2010, [23]) contains about 516,000 manually annotated and reviewed protein sequences in the Swiss-Prot section and more than 10 millions of automatically annotated and not reviewed protein sequences in the TrEMBL section. Hence, being able to retrieve information about new protein sequences is a critical problem.

In this context, automatic tools allowing to provide such information are of big interest. The most common way to perform such a search is to identify patterns specific from a given function for example and to design a prediction method. Information taken into account can consist in different levels: only sequence [2,21], sequence and structure [10,16], only structure [15,12] or use of more general classifications (SCOP [22], GO [6],...). In this paper, the objective is to design a prediction method based on sequence only in order to provide information for only sequenced proteins.

However, sequence-based methods are likely to be limited with regards to structure-based ones as structure is known to be better conserved than sequence [4]. Hence, the proposed method will use a structure-based middle step to identify interesting structural patterns. This step is based upon a Hidden-Markov model structural alphabet (HMM-SA) [3] which proposes a discretized conformation space allowing to accurately describe all possible four residue conformations using 27 prototypes called *structural letters* (SL). This alphabet is known to provide a good description of all secondary structures and especially of loops [19]. This is particularly important as loops are very often implied in interactions [20,1]. In this work, stress is laid on patterns of interest found in loops. Those patterns will be defined here as four SL words encoding seven amino-acid residues. This length has been chosen to obtain satisfying representativities [18]. However, the prediction method is independent on the pattern length and could be applied to any identified pattern.

This paper will focus on the prediction of a pattern (once it has been identified in any way) directly from sequence. One example of pattern identification method will be really briefly discussed. Our prediction method will be divided into two steps: first, each four amino-acid residue will be assigned a profile of possi-

ble structural letters thanks to boolean trees aiming at extracting sequence information. Then, the prediction of words of successive SLs is then assembled through a Hidden Markov Model in order to compute a score for the probability of finding a given functional pattern behind the considered sequence. The method will be applied to patterns identified to be related to ATP- and SAH/SAM-binding sites and to specific turns.

Fig. 1. Illustration of the way of fragment corresponding to seven AA ($AA_1, \ldots, AA_7$) is encoded into a four SL word (SL_1, SL_2, SL_3, SL_4).

2 Material and Methods

2.1 The initial data

We use 16,995 loops extracted from Protein Data Base (PDB) with at most 25% of sequence identity. This sequence identity rate aims at avoiding any bias in the learning step. The length of loops ranges from 1 to 1,261 SL (hence from 4 to 1,264 AA) with an average of 116 and a standard deviation of 129. They are extracted from 7,778 different proteins.

2.2 The transformed: encoding through HMM-SA

As introduced previously, HMM-SA [3] aims at discretizing the conformational space of four-residue fragments into 27 structural states called structural letters (SLs). It is based on a hidden-Markov modelling allowing to take into account dependencies between successive letters. Obtained SLs can be divided into groups: four SLs particularly describe conformation of α-helices (namely a, A, V and W), five SLs describe β-sheets (L, M, N, T and X) and the remaining 18 letters allow the characterization of loops. This is particularly interesting as loop variability is very important.
The Markovian aspect of the model is particularly interesting to study dependencies between successive fragments. Indeed, it can be shown that some transitions between certain SLs are favoured: after a given SL, some SLs will be more probably found than other ones. It can be really informative in a prediction objective to take into account such dependencies.
From this alphabet, it is possible to encode any 3D structure into a 1D string only containing SLs. In this goal, Viterbi or forward/backward algorithms can be used to find the most probable sequence of SLs according to the observed structure. This has been done on the dataset described in section 2.1. In the following sections, focus will be put on four SL words, as illustrated in Fig. 1.

From now on, the goal is to be able to model the link between AA and SL. Indeed, this link is not obvious at all. It is impossible to find a perfect application from the set of four AA sequences ($20^4 = 1.6 \times 10^5$ possibilities) onto the 27 possible SLs. Indeed, the same AA sequence can be encoded into different SLs and the same SL can be obtained from different four-AA sequences.

2.3 First step: from four amino-acids to one structural letter

The goal of the first step is to build classifiers aiming at finding which AA sequence characteristics are the most relevant to distinguish the different SLs. The structure of data is quite particular: the combination of four nominal qualitative variables taken from the same 20 cardinality alphabet are used to predict another nominal qualitative variable with 27 classes.
In order to simplify the problem, an usual one-versus-one classification process will be used. In this way, each classifier has to deal with a simpler problem, optimizing the expectation to finally obtain satisfying predictions. Hence, a classifier will be built for each pair of SLs leading to the optimization of 351 ($(27 \times 26)/2$ classifiers).

2.3.1 Structure of a classifier In order to find robust information about the complex relationship between AA sequence and SLs and to avoid any kind of overfitting, a very simple use of information is proposed. Each classifier consists in a series of binary questions about the presence or absence of one AA at one position in the sequence (let us remind that four-residue sequences are considered for one SL). Then, those questions are combined thanks to AND/OR operators. Actually, a global binary question is obtained through this combination allowing to classify the observations into two groups according to the obtained answer (YES/NO). It is really appropriate to illustrate such a classifier through a tree-like representation. For

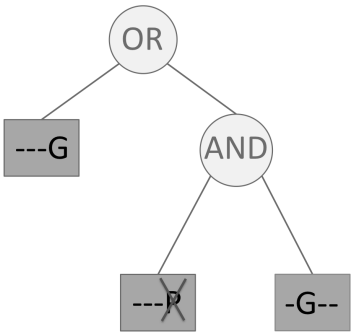


Fig. 2. Example of classifier used to discriminate between two letters A and B: if there is *G in second position AND no P in fourth position* OR a G in fourth position then the sequence is affected to A else to B.

instance, Fig. 2 gives an example. Contrary to classical decision trees, this tree has to be read from leaves to root: by sequentially answering to each leaf question (there is or there is no such letter at such position) and combining the answers through the AND/OR operators contained in nodes, a global yes/no answer is obtained allowing to affect the AA sequence to one of the two SLs compared through this classifier.
A jury of 351 *experts* is finally obtained providing 351 votes concerning the 27 SLs. Hence, a kind of profile in SLs is obtained for a four-residue fragment.

2.3.2 Scoring a classifier In order to choose an appropriate classifier for each pair of SLs, each classifier must be associated with an objective value, called *fitness value*, which quantifies its quality with regards to the classification problem. This value consists in three parts: a term associated to the entropy gain (directly linked with discrimination ability), a term to the tree complexity and a term to the representativeness of the obtained decision rule.
The entropy (we refer to Shannon entropy [5]) is defined here in the context of information theory. If one considers several observations of a random variable, the maximum entropy is obtained if all its possible values are equiprobable.
In our context, the global entropy associated with a sample containing the observations of two SLs i and j ($i \neq j$ and $(i,j) \in \{1,2,\ldots,27\}^2$) can be defined as:

$$\begin{aligned} H(i,j) &= -\left(p_{ij}^{(i)} \log(p_{ij}^{(i)}) + p_{ij}^{(j)} \log(p_{ij}^{(j)})\right) \\ &= -\left(p_{ij}^{(i)} \log(p_{ij}^{(i)}) + (1-p_{ij}^{(i)}) \log(1-p_{ij}^{(i)})\right), \end{aligned}$$

where $p_{ij}^{(k)}$ is the proportion of SL k ($k \in \{i,j\}$) in the sample containing all the observations of SL i and all of the SL j ($j \in \{1,2,\ldots,27\}$) and no other SL. Then, a perfect classifier would divide such a sample into two subsamples, each one containing all the observations of one SL. This situation would represent the best possible entropy gain for these two SLs. More formally, the entropy gain due to the classifier aiming at classifying SLs i and j can be defined as follows:

$$\begin{aligned} G(i,j) = H(i,j) \\ &+\pi_1 \left(p_{ij,1}^{(i)} \log(p_{ij,1}^{(i)}) + p_{ij,1}^{(j)} \log(p_{ij,1}^{(j)})\right) \\ &+\pi_2 \left(p_{ij,2}^{(i)} \log(p_{ij,2}^{(i)}) + p_{ij,2}^{(j)} \log(p_{ij,2}^{(j)})\right), \end{aligned}$$

where $p_{ij,l}^{(i)}$ is the proportion of the SL i in the l-th ($l \in \{1,2\}$) subsample of the sample containing all the observations of SL i and all of the SL j provided by this classifier, π_k ($k \in \{1,2\}$) is the proportion of SLs contained in subsample k (both i and j are taken into account together). Thus, the entropy gain is the difference between the global entropy and the weighted sum of entropies of the two subsamples.
Then, in order to take into account the parsimony of the model, penalizing the complexity of the tree is a way to avoid overfitting and loss of generalizability. We chose to take the number of leaves of a tree as a quantification of its complexity. In order to normalize the variations of the complexity term in the fitness function, the following term is used:

$$penal_1 = \frac{2^{D-1} - nbf}{2(2^{D-2}-1)},$$

where nbf is the number of leaves in the considered tree and D is the maximum authorized depth of any tree.
Finally, another problem may occur which would not be detected by the former two terms (entropy gain and first penalty). Hence, if the classifier makes a relatively small subsample but which contains a very important proportion of only one of the SLs, its entropy will be really satisfying whereas it is not representative of either of the two SLs. This kind of behaviour will be penalized by the following term:

$$penal_2 = \left| \frac{p_{ij,1}^{(i)}}{p_{ij}^{(i)}} - \frac{p_{ij,1}^{(j)}}{p_{ij}^{(j)}} \right|.$$

Indeed, it measures the difference between the proportion of SL i and of SL j found in the 1^{st} subsample. The variation of $penal_2$ is the same as the variation of the same term for the 2nd subsample. As the objective is to obtain two subgroups having a high proportion of the instances of one SL and a small proportion of the other one, the higher this term, the better the classifier.
Finally, the global fitness function can be expressed as follows:

$$fit(i,j) = G(i,j) + \alpha(penal_1 + penal_2),$$

where α allows to balance the entropy gain term and the penalizations. In our applications, we chose $\alpha = 0.05$. It is chosen as the quantification of how much entropy gain we are willing to loose to be able to delete one leaf in the tree.

2.3.3 Optimization of each classifier The kind of decision rule chosen to distinguish between two SLs naturally leads to the use of genetic programming (GP) [9,11] to optimize each *tree*. Indeed, the cardinality of the set of all possible trees is really huge: it is easy to see that it is not possible to exhaustively explore the whole solution space. That is why, a heuristic method has to be used.
GP is a symbolic approach to computer programs induction. It is a kind of genetic algorithm [8,17] where potential solutions are programs defined on a landscape determined by the objective task. In our context, a program will be a classifier (that is to say a tree). Thus, the GP will allow the evolution of a population of potential classifiers through the use mutation and cross-over.

2.4 Second step: looking for one specific functional pattern

Due to the complexity of the prediction problem (impossibility to build an easy bijection between AAs and SLs), the first step cannot be sufficient to answer the problem. Hence, some SLs are easy to discriminate through their sequence. For example, SLs B and M are very well discriminated through their classifier: one out of the two subgroups obtained after applying the classifier contains 3.2% of the B SLs and 98.0% of the M. On the other hand, SLs a and M are particularly difficult to distinguish through their sequence: one out of the two subgroups obtained after applying the corresponding classifier contains all the SLs a and 80.6% of SLs M, which is a very poor classification. Hence, further information has to be taken into account to be able to make decisions about a four-SL word. In this context, a particularly interesting knowledge is about dependencies between successive letters. It is the goal of the second step.
The aim of this step is to decide, given the results of the first step for four consecutive SLs and through a scoring function, if the conformation adopted by the considered seven residue fragment is likely to be encoded by a given four SL word identified to be linked to a functional pattern.

2.4.1 One way of identifying patterns As previously mentioned, pattern identification is not the topic of this paper but here is an exemple of how to find such motifs. First, after encoding step, loop four SL words are systematically extracted in a given non-redundant dataset. Then, the over-representation of each word is quantified [18] to identify words which are likely to be of interest. Words with important over-representation are considered as interesting candidates to be functional patterns.
Finally, candidate locations are crossed to Swiss-Prot annotations to see if they are specific of any functional indication. Examples are given in section 3. It is important to notice that among identified patterns only the ones showing sequence specificities will be likely to be predicted directly from sequences.

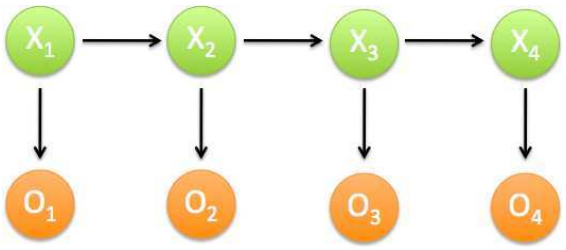


Fig. 3. Structure of the HMM used to model the relationship between first step outputs and *true* SLs for a seven residue fragment: $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4)$ are the *true* SLs and $\mathbf{O}_i = (o_i^1, o_i^2, \ldots, o_i^{351})$ is the vector of votes obtained from step 1 for the four AA fragment encoded by X_i.

2.4.2 Pattern modelling and application to prediction As emphasized earlier, a real dependency exists between successive SLs, especially because of overlaps. Hence, this dependency can be favourably used to build a model. A Hidden-Markov Model (HMM) has been chosen to model the link between first step outputs and a given 4 SL pattern. This HMM is described in Fig. 3. In this model, hidden states are the *true* SLs while observed states are outputs of step 1 for the corresponding sequence. Arrows between $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$ symbolizes the dependency between successive letters called *transition probabilities* in HMM context and arrows between $\mathbf{X}_i$ and $\mathbf{O}_i$ represent the link between true SLs and step 1 outputs, namely the *output probabilities*.

Thanks to this model, the objective of the second step is to compute the probability of the four true SLs being the target functional pattern given the step 1 outputs for four successive (and overlapping) four AA fragments. Hence, we have to compute

$$P(X_{1:4}|O_{1:4}) = P(X_1, X_2, X_3, X_4|O_1, O_2, O_3, O_4). \quad (1)$$

High values of this probability will indicate a strong assumption that the considered fragment is likely to be encoded into the identified pattern and then to have the target function.
According to the chosen model,

$$P(X_{1:4}|O_{1:4}) = P(X_1|O_1)\prod_{i=2}^{4} P(X_i|X_{i-1})P(X_i|O_i).$$

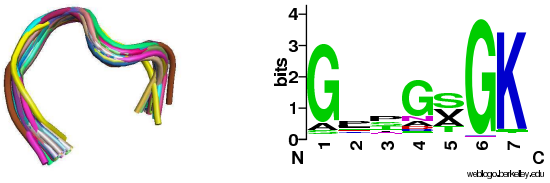
Now, $P(X_i|O_i)$ has to be computed. Assuming that the results of different trees are independent,

$$P(X_i|O_i) = P(X_i|o_i^1, o_i^2, \ldots, o_i^{351}) = \prod_{j=1}^{351} P(X_i|o_i^j).$$

This assumption is wrong for some comparisons (especially comparisons implying a common SL which is well predicted) but most of pairs of comparisons can be considered as independent (results not shown).
Then, by Bayes theorem,

$$P(X_i|o_i^j) = \frac{P(o_i^j|X_i)P(X_i)}{P(o_i^j|X_i)P(X_i) + P(o_i^j|\overline{X_i})P(\overline{X_i})}.$$

Finally, $P(X_i)$, $P(o_i^j|X_i)$ and $P(X_i|X_{i-1})$ are estimated on the dataset.



(a) (b)

Fig. 4. (a) Representation of several fragments encoded into $YUOD$. (b) Weblogo of the AA sequences encoded into $YUOD$.

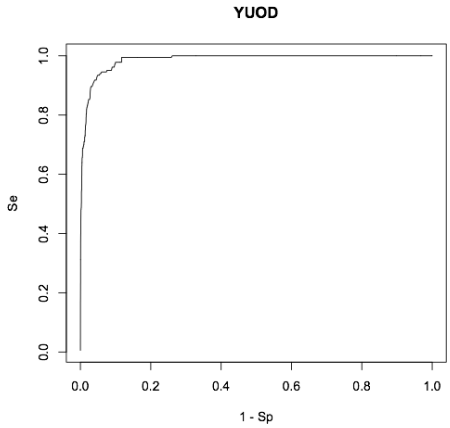


Fig. 5. ROC curve associated with the probability of having $YUOD$ for a given seven AA fragment (Se=sensitivity, Sp=specificity).

3 Applications

3.1 Prediction of an ATP-binding site specific motif

Previous studies (as described in section 2.4.1) have shown that fragments encoded into the four SLs $YUOD$ (see Fig. 4(a)) are very often associated to ATP/GTP binding sites. Indeed, on our database, 95% of fragments encoded into $YUOD$ are associated to this function in SwissProt. Hence, being able to predict the encoding into YUOD is really useful to predict this function for a new sequence.
The superposition of several fragments encoded into $YUOD$ is shown in Fig. 4. Moreover, this structural word has a high sequence specificity as shown in Fig. 4, especially positions 1, 6 and 7. Thus, this structural word is a very good candidate for our approach.

In our dataset, $YUOD$ can be found 183 times in 181 proteins (two proteins contain two occurrences). The model is applied on the whole proteins to study the ability of the computed probability (Eq. 1) to discriminate between $YUOD$ and $\overline{YUOD}$ (*not YUOD*). The ROC curve associated to the logarithm of this probability is shown in Fig. 5. It displays the sensitivity (ability to retrieve $YUOD$) and specificity (ability to recognize $\overline{YUOD}$) according to the probability threshold chosen to split the words into $YUOD$ and $\overline{YUOD}$. The AUC (area under curve) associated to this ROC curve is 0.9866. Hence, the computed probability is really efficient to identify $YUOD$ among all other words. Indeed, such a discrimination quality is particularly valuable because of the ratio between the two classes: $YUOD$ only represents 0.52% of studied words. Then, according to the application requirements, several thresholds can be defined providing different balances between sensitivity and specificity. Some interesting threshold values and their corresponding parameters are enclosed in Tab. 1. Very high values of specificity have been chosen, indeed the $\overline{YUOD}$ class is really large and then only 1% of false positive ($\overline{YUOD}$ predicted as $YUOD$) can be a large number when applied to big proteins or to several proteins.

An example of $YUOD$ detection is given in Fig. 6. It concerns the Circadian clock protein kinase kaiC, chain A (pdb ID: 2gbl_A). It originally contains two true $YUOD$ occurrences and four have been predicted through our model. Two out of the four positives (numbers 1 and 2) are exactly located at co-

Threshold	-4829	-4805	-4732
Specificity	90.02	95.07	99.00
Sensitivity	97.81	93.44	69.95

Tab. 1. Sensitivity and specificity obtained for the identification of $YUOD$ according to the chosen log(probability) threshold.

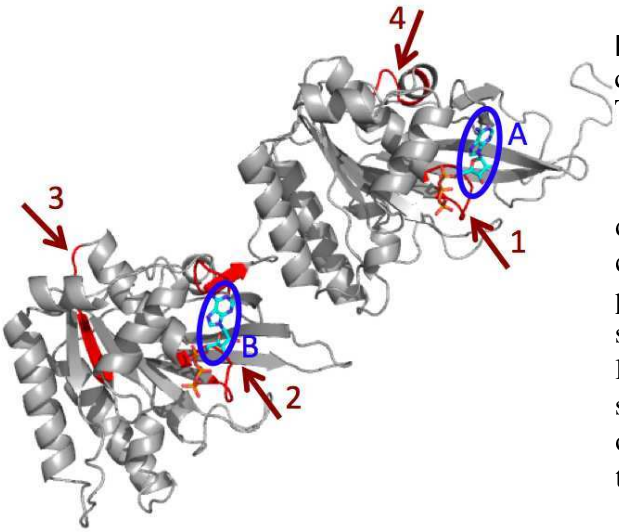


Fig. 6. 3D representation of 2gbl_A co-crystallized with two ATP molecules (indicated by lettered circles). The fragments identified as $YUOD$ are indicated with numbered arrows.

crystallized ATP binding sites (A and B). Moreover, among the two false positives, number 3 adopts a 3D conformation which is really close to the one observed at ATP binding sites. This example demonstrates the difficulty of evaluating a prediction method for annotations. The evaluation of true positive and false negative can be really precise when dealing with manually annotated and reviewed databases such as SwissProt but false positives may be true positive that have not yet been experimentally verified. It is impossible to make a decision in this case.

3.2 Prediction of a SAH/SAM-binding site specific motif

S-adenosyl-methionine (SAH/SAM) and S-adenosyl-homocysteine are molecules associated to some methylation processes and are particularly studied in the context of antiviral drugs research. It is then interesting to be able to predict their binding to proteins. The four-SL word $RUDO$ has been identified to be most of time associated to SAH/SAM in SwissProt. Moreover, it has a certain sequence specificity (results not shown).
In our dataset, $RUDO$ is found 39 times in 39 different proteins. The AUC associated to the ROC curve corresponding to the log(probability) computed by our method is 0.9606. The specificity and sensitivity obtained with different thresholds for the log(probability) are given in Tab. 2. Thus, results are satisfying and allow to recover more than two thirds of the $RUDO$ motifs without wrongly assigning more than 1% of the other words.

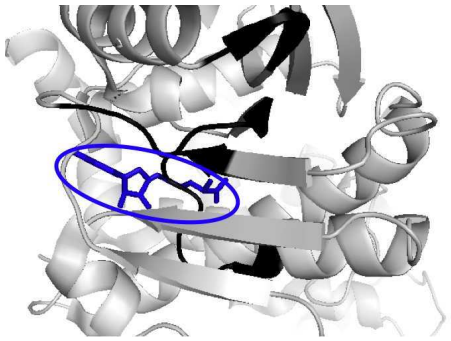


Fig. 7. 3D representation of 1fp1 (light grey cartoons) co-crystallized with a SAH molecule (indicated by a circle). The fragments identified as $RUDO$ are black-coloured.

Threshold	-4903	-4806	-4712
Specificity	90.00	95.00	99.00
Sensitivity	87.18	84.62	69.23

Tab. 2. Sensitivity and specificity obtained for the identification of $RUDO$ according to the chosen log(probability) threshold.

An illustration can be found in Fig. 7. It concerns isoloquiritigenin 2'-O-methyltransferase (PDB ID: 1fp1) which was here co-crystallized with a SAH molecules. Four words were predicted as $RUDO$ with a threshold of -4712 whereas only one has been encoded as $RUDO$. However, by looking of the 3D conformation, it appears that all four identified fragments are really closed to the ligand. Thus, the method which uses the HMM-SA as a tool to discover patterns, is not limited to the fragments being strictly encoded but is also able to discover fragments with close structures as only sequence is finally taken into account. Hence, fragments which are likely to adopt a $RUDO$-like conformation will be as well identified by the method.

3.3 Prediction of a specific turn

The prediction of turns is also of special interest in protein study [7]. The four-SL word $HBDS$ can be linked to turns: the corresponding fragment conformations are shown in Fig. 8. This is a frequent word, in our database, it was found 1633 times in

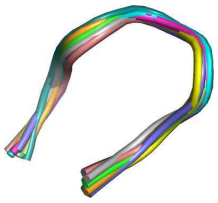


Fig. 8. Representation of several fragments encoded into $HBDS$.

1363 different proteins (there are one to six occurrences in those proteins). The AUC associated to the prediction of $HBDS$ is 0.9359. Tab. 3 indicates the specificities and sensitivities associated to different log(probability) values. The results are a bit less efficient than previous ones (due to a lower sequence specificity) but enable to locate 85% of those turns with a specificity of 90% (knowing this specificity is likely to be underestimated because of close fragments which have not been strictly encoded into HBDS).

Threshold	-4013	-3844	-3777
Specificity	90.17	95.11	98.88
Sensitivity	84.71	71.07	28.93

Tab. 3. Sensitivity and specificity obtained for the identification of $HBDS$ according to the chosen log(probability) threshold.

4 Conclusion and perspectives

The automatic annotation of simply sequenced proteins is a very important task in the present context of high-throughput sequencing programs. The method proposed in this paper is based on the identification of patterns of interest directly on structures through HMM-SA. The **input data** of the described method are **only sequences** and as a consequence, only patterns having sequence specificities will be likely to be handled with this method. But for this kind of motifs, the method is really powerful. One method ([13]) has already been proposed to predict the 3D structure of small peptides through HMM-SA but the motif-oriented aspect of the method proposed here makes it much more precise and time efficient.

As much information as possible is extracted from data. The dependance between AA sequences and 3D structure is learnt in the first step through the use of HMM-SA. Then, the second step takes advantage of two different sources by building a hidden Markov model. Firstly, the strength of dependance between AAs and SLs is quantified and used through observation probabilities: some observations will be really trusted (when a strong link has been found in the first step) whereas others will be considered with care as less reliable. Secondly, the dependance between successive SLs (some SLs favourably follow other ones) is also taken into consideration by the computation of transition probabilities. Finally, a really complete model is obtained by the addition of both steps.

Moreover, as HMM-SA is only an intermediate between sequence and function (or any other interesting pattern), the method, as shown in some illustrations, is able to identify fragments as close to the target word even if this fragment would not be encoded into the exact target word. Hence, relying on sequences is a good way to overcome some cases of flexibility: in the crystallization conditions, the fragment has not been found in the strict conformation associated to the target word, but its sequence specifities can be recognized by the prediction method. Eventually, HMM-SA encoding and the proposed prediction method are interestingly complementing each other in the prediction of patterns of interest.

Furthermore, the important adaptability of the prediction method is of big interest. Indeed, in this paper we focused on pattern which had been identified directly through HMM-SA but it is completely possible to identify 3D motifs as interesting for any reason, to encode it into HMM-SA and to build the model on the obtained word. Let us recall here that the size of considered fragments is not limited. Earlier, only seven-residue fragments have been considered but any length would be possible. Furthermore, as illustrated through the three examples, the size of the learning dataset can be really variable (from 35 to 1633 occurrences of the pattern) as the model is always the same. The only variable parameter is the log(probability) threshold. However, preliminary studies seem to indicate that this threshold depends on the strength of the sequence specificity of the structure. Hence, further work could be able to set this threshold directly from the quantification of this dependance.

The limits of the proposed prediction method are interlocked with its strengths. First of all, as previously indicated, only pattern with sequence specificities can be predicted. Moreover, a 1D intermediate is necessary. HMM-SA has been used because of its very interesting abilities of precise description especially for loops, but the same methodology could be applied on other types of alphabets. Finally, the method is bounded by the function specificity of the pattern. Indeed, a function might be associated to different patterns. Thus, our method is able to predict one type of realization of a given function at a time. Of course, it is completely possible to learn several

patterns linked to a function and to give a global prediction for all of them. But for the moment, this limit prevents us to compare with prediction methods for specific function (such as [2]) encoded through different patterns. This should be quickly possible by the identification of new patterns which is in progress.

References

[1] S. Ansari and V. Helms, Statistical analysis of predominantly transient protein-protein interfaces. *Proteins*, 61:344-355, 2005.

[2] H.R. Ansari and G.P.S. Raghava, Identification of NAD interaction residues in proteins. *BMC Bioinformatics*, 11(160), 2010.

[3] A.-C. Camproux, R. Gautier and P. Tuffery, A hidden Markov Model derived structural alphabet for proteins. *Journal of molecular biology*, 339(3):591-605, 2004.

[4] C. Chothia, J. Gough, C. Vogel and S.A. Teichmann, Evolution of protein repertoire. *Science*, 300(5626):1701-1703, 2003.

[5] T.M. Cover, and J.A. Thomas, *Elements of information theory*, 2nd edition, Wiley, New York, 2006.

[6] J. Espadaler, E. Querol, F.X. Aviles and B. Oliva, Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics*, 22(18):2237-2243, 2006.

[7] P.F.J. Fuchs, A.J.P. Alix, High accuracy prediction of β-turns and their types using propensities and multiple alignments. *Proteins*, 59(4):828-839.

[8] D.E. Goldberg *Genetic Algorithms: Search, Optimization and Machine Learning*, Addison-Wesley, New York, 1989.

[9] J.R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, 1992.

[10] I. Halperin, D.S. Glazer, S. Wu and R.B. Altman, The FEATURE framework for protein function annotation: modeling new functions, improving performance and extending to novel applications. *BMC genomics*, 9(Sup 2):S2, 2008.

[11] W.B. Langdon and R. Poli, *Foundations of Genetic Programming*, Springer-Verlag, 2002.

[12] K. Manikandan, D. Pal, S. Ramakumar, N.E. Brener, S.S. Iyengar and G. Seetharaman, Functionally important segments in proteins dissected using Gene Ontology and geometric clustering of peptide framgents. *Genome Biology*, 9(3), 2008.

[13] J. Maupetit, P. Derreumaux and P. Tuffery, PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic Acids Research*, 37, 2009.

[14] A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536-540, 1995.

[15] B.J. Polacco and P.C. Babbitt, Automated discovery of 3D motifs for protein function annotation. *Bioinformatics*, 22(6):723-730, 2006.

[16] G. Pugalenthi, K.K. Kumar, P.N. Suganthan and R. Gangal, Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. *Biochemical and Biophysical Research Communications*, 367(3):630-634; 2008.

[17] C.R. Reeves d J.E. Rowe *Genetic algorithms - Principles and perspectives, A guide to GA theory*, Kluwer Academic Publishers, London, 2003.

[18] L. Regad, J. Martin and A.-C. Camproux, Identification of non random motifs in loops using a structural alphabet. *Proceedings of IEEE Symposium on computational intelligence in bioinformatics and computational*, 92-100, 2006.

[19] L. Regad, J. Martin, G. Nuel and A.-C. Camproux, Mining protein loops using a structural alphabet and statistical exceptionality. *BMC Bioinformatics*, 11, 2010.

[20] M. Saraste, P.R. Sibbald and A. Wittinghofer, The P-loop: a common motif in ATP- and GTP-binding proteins. *Trends in Biochemical Science*, 15:430-434, 1990.

[21] C.J.A. Sigrist, L. Cerutti, E. de Castro, P.S. Kangendijk-Genevaux, V. Bulliard, A. Bairoch and N. Hulo, PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Research*, 38:D161-D166, 2010.

[22] A.V. Tendulkar, M. Krallinger, V. de la Torre, G. Lopez, P.P.Wangikar and A. Valencia, FragKB: structural and literature annotation resource of conserved peptide fragments and residues. *PLoS one*, 5(3), 2010.

[23] C.H. Wu, R. Apweiler, A. Bairoch, D.A. Natale, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi and B. Suzek, The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, 34:D187-D191, 2006.

Scalability of large-scale protein domain inference

Clément REZVOY[1], Daniel KAHN[2] and Frédéric VIVIEN[3]

[1] ENS Lyon, Université de Lyon, LIP, UMR 5668, ENS Lyon - CNRS - INRIA - UCBL, Lyon, FRANCE
Clement.Rezvoy@ens-lyon.fr
[2] Université de Lyon, Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, UCBL - CNRS, INRA, Villeurbanne, FRANCE
Daniel.Kahn@inrialpes.fr
[3] INRIA, Université de Lyon, LIP, UMR 5668, ENS Lyon - CNRS - INRIA - UCBL, Lyon, FRANCE
Frederic.Vivien@inria.fr

Abstract *The exponential growth of sequence databases challenges the automatic inference of protein domain families. This growth prohibits the use of traditionally sequential algorithms like* MKDOM2*, the algorithm that was used to construct the* PRODOM *database. We present a distributed algorithm for protein domain inference,* MPI_MKDOM2*. This algorithm greatly speeds the processing of large databases, therefore enabling the construction of new versions of* PRODOM*, while preserving the structure of protein domain families built by* MKDOM2*.*

Keywords Sequence clustering, protein domains, distributed computing.

1 Introduction

The PRODOM database is a repository of protein domain families inferred automatically from homologies between the protein sequences of the Uniprot database. Since 1999, PRODOM has been built using MKDOM2 [2], a sequential algorithm of quadratic complexity. Because of its sequential nature, MKDOM2 can not keep up with the exponential increase of Uniprot over the years to the point that it is no longer possible to envision a new release of PRODOM built using the same method: given past records, running MKDOM2 on release 11.1 of Uniprot would last more than 16 years.

MKDOM2 is based on a simple assumption: the shortest sequence of the dataset is the most likely single-domain sequence of the lot. MKDOM2 creates a first family containing segments being found homologous to the shortest sequence of the set by PSI-BLAST [1]. The newly defined family is then removed from the dataset and MKDOM2 iterates until complete exhaustion of the dataset. MKDOM2 is sequential by definition since families must be created one after the other to follow the heuristic. The PSI-BLAST searches represent most of the total computation and yet each individual search is in average too short to yield good performance when distributed. To create an efficient distributed algorithm we relaxed the original heuristic to allow several queries to be ran at once. This document presents the new parallel algorithm MPI_MKDOM2 and outlines its performance as well as the stability of sequence clustering with respect to that of the original sequential algorithm.

2 Parallel execution strategy

MPI_MKDOM2 is a master-worker algorithm were the master distributes the query sequences to the workers and gather the results afterwards. Running several queries in parallel leads to several problems that has been addressed in MPI_MKDOM2.

Conflict avoidance. To ensure that results are similar to those of mkdom2 with no overlapping families, the results of queries are validated sequentially. If a query result overlaps with a previously defined family, the result is not taken into account. Conflicts between query results are wasteful in resource. Either the results correspond to a domain family wrongfully computed twice, or the results marginally overlap with a previous family and will have to be recomputed once the database has been updated. In order to avoid pathological cases of overlaps, we rely on the results of an all-against-all BLAST search of the database. Sequences found sharing homology within this first computation will not be processed in parallel. PSI-BLAST being less stringent than a BLAST this pre-computation cannot guarantee the absence of conflicts. It is however sufficient to maintain occurrences of such conflicts at a level that does not impede parallel performance.

Absorption of variations of query running time. Instances of PSI-BLAST have large variations in running time. To level these variations, each worker is alloted several queries at once. The master will start selecting new queries once one of the worker has computed half of its batch. Moreover workers are desynchronized, the first worker to finish does not have to wait for the last worker to carry on with new queries.

3 Experimental evaluation

In order to assess the performance of the parallelization, MPI_MKDOM2 was used to process a database (DB) of 556,964 sequences (PRODOM 2003.1 input data) and a database of 69,621 sequences (DB/8) corresponding to a randomly chosen eighth of DB.

Parallel efficiency. Tab. 1 shows the time required to process DB/8 with an increasing number of workers. MPI_MKDOM2 manages to reach a speedup (acceleration factor with respect to the 1–worker case) of 25 while maintaining a high efficiency (total compute time divided by the total compute time of the 1–worker case). After a quick initial increase the speedup reaches a plateau near 33 between 39 and 79 workers, before collapsing. However, the larger the database, the larger the speedup. For instance, the processing time of DB/8 is only divided by 1.11 going from 39 to 79 workers while that of DB is divided by 1.72 .

Number of workers	Wall-clock duration	Total compute time	Speedup	Efficiency
1	14h20'32"	14h20'32"		
2	7h09'37"	14h19'14"	2.00	1.00
7	2h12'50"	15h29'54"	6.48	0.93
31	0h33'32"	17h19'51"	25.65	0.83
39	0h29'00"	18h51'32"	29.66	0.76
63	0h26'17"	27h36'41"	32.72	0.52
79	0h26'10"	34h28'24"	32.87	0.42
127	0h40'28"	85h40'26"	21.26	0.17

Tab. 1. Summary of parallel performance results for the processing of the DB/8 database.

Conflict avoidance efficiency. Using the pre-computed all-against-all result maintains the occurrences of conflicts under 6 % of the total number of queries processed. Using conflict prevention we obtain a conflict ratio of 1.59 % and a speedup of almost 30 for 39 workers. In comparison running MPI_MKDOM2 on the same database with as many workers but without conflict prevention leads to more than 25 % of conflicts for a speedup of only 2.

Number of workers	1	2	7	31	39	63	79	127
% of conflicts	0.37	0.47	0.70	1.35	1.59	3.13	3.75	5.11

Tab. 2. Percentage of queries leading to conflicts as a function of the number of workers, for the processing of DB/8.

Result stability. To assess the proximity between results we measured the Wallace1 (W1) index [3] between the results of MKDOM2 and MPI_MKDOM2 with a variable number of workers (Fig. 1). The W1 index measures the probability that a pair of residues clustered in the same family by MKDOM2 are also grouped by MPI_MKDOM2. To compute global indices, W1 was calculated for all families inferred by MKDOM2 and weighed by the numbers of residues

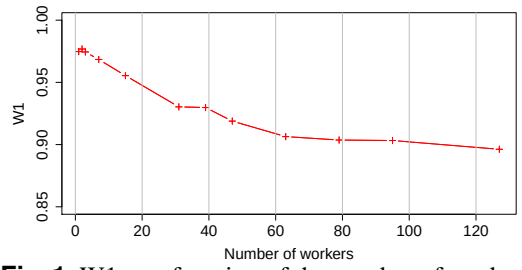


Fig. 1. W1 as a function of the number of workers.

per family. The global W1 index decreases following the increase in the number of workers used. Yet it manages to stay above 0.90 as long as fewer than 100 workers are used. Looking at individual W1 index for each family shows that this degradation does not depend on the size of families and more than 80 % of families have a W1 index above 95 %.

4 Conclusion

The distributed algorithm presented here is able to provide reasonable speed-ups while retaining the structure of the protein domain families built buy the sequential algorithm. Even when achieving maximal speedup the clustering stays consistent with the sequential result. The speedup being dependent of the size of the database processed, this new algorithm will provide a mean to efficiently construct new releases of PRODOM while staying consistent with the original sequential heuristic.

Acknowledgements

The PRODOM project was supported by the FP6 EMBRACE Network of Excellence and the FP7 IMPACT Research Infrastructure Programme. Experiments presented in this paper were carried out using the Grid'5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS and RENATER among others (see https://www.grid5000.fr). This work was performed using HPC resources from GENCI-CINES (Grant 2010-c2010076425).

References

[1] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

[2] J. Gouzy, F. Corpet, and D. Kahn. Whole genome protein domain analysis using a new method for domain clustering. *Comput Chem*, 23(3-4):333–40, 1999.

[3] D.L. Wallace. A method for comparing two hierarchical clusterings: comment. *Journal of the American Statistical Association*, 78(383):569–576, 1983.

Counting RNA pseudoknotted structures

Cédric SAULE[1,4], Mireille RÉGNIER[1,2,4], Jean-Marc STEYAERT[2,4] and Alain DENISE[1,3,4]

[1] LRI, UMR8623 CNRS, Bât 490, Université Paris-Sud 11, 91405 Orsay, Cedex, France
{saule,denise}@lri.fr
[2] LIX, UMR X-CNRS, Ecole Polytechnique, 91128 Palaiseau, Cedex, France
steyaert@lix.polytechnique.fr,mireille.regnier@inria.fr
[3] IGM, UMR8621 CNRS, Bât 400, Université Paris-Sud 11, 91405 Orsay, Cedex, France
denise@lri.fr
[4] INRIA Saclay, Parc Orsay Université, 4 rue Jacques Monod, 91893 Orsay, Cedex, France
steyaert@lix.polytechnique.fr,mireille.regnier@inria.fr,{saule,denise}@lri.fr

Abstract *In 2004, Condon and coauthors gave a hierarchical classification of exact RNA structure prediction algorithms according to the generality of structure classes that they handle. We complete this classification by adding two recent prediction algorithms. More importantly, we precisely quantify the hierarchy by giving closed or asymptotic formulas for the theoretical number of structures of given size n in all the classes but one. This allows to assess the tradeoff between the expressiveness and the computational complexity of RNA structure prediction algorithms.*

Keywords RNA structures, pseudoknots, enumeration, asymptotics, algorithmic complexity.

Enumération de structures ARN avec pseudonoeuds

Résumé *En 2004, Condon et ses coauteurs ont défini une classification des algorithmes exacts de prédiction de structure d'ARN, selon le degré de généralité des classes de structures qu'ils sont capables de prédire. Nous complétons cette classification en y ajoutant deux algorithmes récents. Chose plus importante, nous quantifions la hiérarchie des algorithmes, en donnant des formules closes ou asymptotiques pour le nombre théorique de structures de taille donnée n dans chacune des classes, sauf une. Ceci fournit un moyen d'évaluer, pour chaque algorithme, le compromis entre son degré de généralité et sa complexité.*

Mots-clefs Structures d'ARN, pseudonoeuds, asymptotique, complexité algorithmique.

En 2004, Condon *et al.* publièrent une classifications des algorithmes exacts de prédiction de structure d'ARN selon le degré de généralité des structures qu'ils peuvent prédire [3]. Ils considérèrent les classes de structures sans pseudonoeuds [6,10] (PKF) et les classes suivantes pour les structures avec pseudonoeuds : Lyngso et Pedersen (L&P) [5] , Dirks et Pierce (D&P) [4] , Akutsu et Uemura (A&U) [1,9] , et Rivas et Eddy (R&E) [8]. Chacune de ces classes représente l'ensemble des structures qui peuvent être solutions d'un algorithme de prédiction de structure. La complexité de ces algorithmes s'étend entre $O(n^3)$ pour les structures sans pseudo-noeuds et $O(n^6)$ pour la classe R&E, pour une séquence de longueur n. Condon *et al.* prouvèrent les inclusions suivantes :

$$PKF \subset L\&P \subset D\&P \subset A\&U \subset R\&E.$$

Notre but est de quantifier ces relations, et par la même occasion d'évaluer le compromis entre complexité en temps et nombre de structures prédictibles en théorie pour chaque algorithme. Pour ce faire, nous donnons des formules asymptotiques pour le nombre de structures de taille n dans chacune des classes. De plus, nous ajoutons à notre étude deux nouvelles classes correspondant à des algorithmes publiés postérieurement à 2004 : la classe R&G pour Reeder et Giegerich [7] et la classe C&C pour Cao et Chen [2]. Les structures que nous considérons sont "épurées", dans le sens où l'on ne considère que les nucléotides appariés. La taille d'une structure est son nombre de paires de bases. Pour énumérer la classe L&P, nous présentons une bijection entre l'ensemble des structures L&P de taille n et l'ensemble des cartes planaires enracinées sans isthme à n arêtes et un ou deux sommets. D'autre part, nous montrons que les classes L&P, D&P, A&U, R&G, C&C peuvent être codées par des langages algébriques non ambigus. A partir d'une grammaire non ambiguë pour chacun des langages, nous obtenons une équation pour la série

génératrice dont nous déduisons en équivalent asymptotique pour le nombre de structures de taille n.

Nous établissons que, à l'exception de la classe L&P dont la formule asymptotique est plus simple, le nombre de structures de taille n est, asymptotiquement, de la forme

$$\frac{\alpha}{2\sqrt{\pi}\, n^{3/2}}\omega^n\,,$$

où α et ω sont deux constantes qui dépendent de la classe considérée. Le tableau suivant présente nos principaux résultats. Nous indiquons par une astérisque les classes qui n'avaient pas été dénombrées jusqu'ici. La classe "Toutes" désigne l'ensemble des structures avec pseudonoeuds, en bijection avec les involutions sans points fixes.

Classe	asympt.	α	ω	Compl.	Remarque
PKF	$\frac{\alpha}{2\sqrt{\pi}\, n^{3/2}}\omega^n$	2	4	$\mathcal{O}(n^3)$	Nombres de Catalan
L&P *	$\frac{1}{2}\omega^n$	-	4	$\mathcal{O}(n^5)$	Formule close
C&C *	$\frac{\alpha}{2\sqrt{\pi}\, n^{3/2}}\omega^n$	0,1707	5,857	$\mathcal{O}(n^6)$	
R&G *	$\frac{\alpha}{2\sqrt{\pi}\, n^{3/2}}\omega^n$	0,1521	6,576	$\mathcal{O}(n^4)$	
D&P *	$\frac{\alpha}{2\sqrt{\pi}\, n^{3/2}}\omega^n$	0,7535	7,314	$\mathcal{O}(n^5)$	
A&U *	$\frac{\alpha}{2\sqrt{\pi}\, n^{3/2}}\omega^n$	0,6585	7,547	$\mathcal{O}(n^5)$	
R&E	ouvert	-	-	$\mathcal{O}(n^6)$	
Toutes	$\sqrt{2}\cdot 2^n\cdot\left(\frac{n}{e}\right)^n$	-	-	NPC	Involutions sans points fixes

Il est visible que, d'un point de vue strictement comptable, la multiplication par un facteur n^2 de la complexité entre PKF et L&P ne se justifie pas par le faible gain en nombre de structures possibles. Le cas est pire encore pour la classe C&C qui présente une complexité plus forte que R&G alors que sa cardinalité est exponentiellement plus faible (on montre d'ailleurs aisément que C&C $\subset$ R&G). En revanche, par exemple, l'augmentation linéaire de complexité entre D&P et R&G nous semble très raisonnable en regard de l'augmentation exponentielle du nombre de structures possibles.

Références

[1] T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104 :45–62, 2000.

[2] S. Cao and S-J Chen. Predicting structured and stabilities for h-type pseudoknots with interhelix loop. *RNA*, 15 :696–706, 2009.

[3] A. Condon, B. Davy, B. Rastegari, S. Zhao, and F. Tarrant. Classifying RNA pseuknotted structures. *Theorical computer science*, 320 :35–50, 2004.

[4] N.A. Dirks, R.M. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J Comput Chem*, 24 :1664–1677, 2003.

[5] R. B. Lyngsø and Pedersen C. N. RNA pseudoknot prediction in energy based models. *Journal of computational biology*, 7 :409–428, 2000.

[6] R. Nussinov, G. Pieczenik, J. R. Griggs, and Kleitman D. J. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35 :68–82, 1978.

[7] J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5 :104, 2004.

[8] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of molecular biology*, 285 :2053–2068, 1999.

[9] Y. Uemura, A. Hasegawa, S. Kobayashi, and T. Yokomori. Tree adjoining grammars for RNA structures prediction. *Theorical computer science*, 210 :277–303, 1999.

[10] M. Zucker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acid Research*, 9 :133–148, 1981.

Computational biology exploration of the enzymatic diversity of an uncharacterised prokaryotic protein family

Adam SMITH, Marcel SALANOUBAT, Jean WEISSENBACH, Claudine MEDIGUE and David VALLENET

CEA, IG, Genoscope, 2 rue Gaston Crémieux CP5702, F-91057 Evry, France,
CNRS, UMR 8030, 2 rue Gaston Crémieux, CP5706, F-91057 Evry, France,
Université d'Evry, F-91057 Evry, France.

{asmith, vallenet}@genoscope.cns.fr

Abstract *An estimated 3107 Pfam protein families (26% of 11912) do not correspond to any known functions, even though occasionnaly, a biochemical breakthrough may lead to the functional characterisation of a part of a family. Under the hypothesis of a moderately conserved enzymatic mechanism or of substrate similarity, the previous discovery becomes a foothold in the family's functional space, that computational biology methods can help exploit. Here, we present such a case, along with the phylogenetic and genomic/metabolic context-based strategies used to help explore the family's functional diversity, and to help guide the biochemical assays required for experimental validation.*

Keywords protein family, enzymatic activity, genomic context.

1 Context

In 2006, a joint work combining a comparative genomics approach with biochemical experiments led to the discovery of the coding gene for a previously orphan enzymatic activity participating in a degradation pathway of lysine [1]. The corresponding protein belonged to a known prokaryotic protein family of unknown function, established by Pfam [2] on the basis of domain conservation. However, not all organisms with proteins from this family shared the other enzymes from the lysine degradation pathway ; we hypothesised that a yet unknown diversity of novel but related enzymatic functions was waiting to be discovered, a family of reactions we have temporally named BKACE for "beta-keto acid cleaving enzyme".

In order to explore this potential functional diversity, we chose to conduct a bioinformatics analysis upon the family's proteins, integrating information from different sources, in order to 1) define functional sub-groups in the family to help build a representative selection of proteins for biochemical testing, and 2) propose one or several potential novel activities/substrates per sub-group.

2 Strategy

Our first objective was allegedly to propose a list of candidate genes for cloning, that hopefully would span the family's potential functional and sequence spaces. In order to do this, we generated several different unsupervised clusterings for the proteins, each based on different data sources, which we then integrated into a final clustering (defining our sub-groups) using an ensemble clustering approach [3].

The set of protein sequences from the BKACE family were first aligned one-to-one using BLASTP. The log-evalues were used as a similarity measure and fed to a complete-linkage algorithm, defining a sequence homology-based clustering.

A multiple sequence alignment (MSA) was made from the set of sequences using MAFFT [4] before manual curation. It was then used to build a bootstrapped phylogenetic tree using quicktree [5], which served as a basis for the manual creation of a phylogenetic clustering. We also used SCI-PHY [6] on the MSA to build another homology-based clustering.

The most original data source used in our strategy was BKACE genomic contexts. Here, we define a genomic context of a BKACE protein as a set of at least two genes co-localised on one genome with a BKACE, that are conserved in at least another genome with a BKACE. Gene gaps of up to three

genes were allowed. Conserved contexts were calculated using the in-lab Syntonizer tool [7,8].

With these genomic contexts, it became possible to define a new measure of similarity between two BKACE proteins. The simplest measure, implemented here, was to count the number of conserved genes, for each pair of BKACE proteins. The resulting similarities were then processed into a graph, to which we applied a spectral clustering algorithm [9], generating a new clustering (hereafter "GC").

Another novel clustering was generated by a 3D protein modelling method, which specifically identified the amino acids present in the protein's active site, and clustered the proteins accordingly [10].

Finally, all these clusterings were combined into a single clustering (hereafter "MegaClustering") using hard ensemble clustering, a statistical approach using partition distance metrics to estimate a partition "as close as possible to all others" [3].

Additionally, the GC clusters formed the basis of another analysis. In order to reduce the space of possible activities of the family, we postulated that all BKACEs should exhibit a reaction mechanism not too dissimilar from the known one. Two sources of information could help guess which similar activities could exist : overall substrate similarity, and metabolic context. "Metabolic context" refers here to the sum of enzymatic activities exhibited by all the genes in the same GC cluster of a studied BKACE. Potential substrates were manually proposed on the basis of these metabolic contexts.

3 Results & Discussion

After manual removal of proteins with aberrant sequence lengths, start/stop codon problems, and misalignment problems, 725 BKACE protein sequences were extracted from the Pfam family, spanning 141 prokaryotic genera. Our MegaClustering approach grouped these proteins into 32 clusters.

Bacterial clone preparation for protein overexpression being a difficult and empiric science (especially given the G-C richness of many BKACE proteins), candidate proteins were chosen manually by the biochemists from each MegaCluster.

Analysis of the pooled metabolic contexts from each of the GC clusters led to the proposition of 5 potential substrates for BKACE activities. Several additional substrates were proposed by our partners on the basis of substrate-similarity searches.

At the time of writing, 54 BKACE proteins (covering 18 (56%) of MegaClusters) have been cloned and tested for the proposed BKACE activities. Many seem to have a wide substrate specificity. It has been manually observed that activity profiles seem to concord with the MegaClustering; once enough proteins have been tested, we will carry out statistical analyses in order to verify this preliminary observation.

4 Conclusion

Our analysis of an uncharacterised protein family has led to a clustering of practical interest for wet-lab experiments. Furthermore, use of genomic/metabolic contextual information (*i.e.*, the functional annotations of neighbouring co-conserved genes) led to the proposal of several potential activities to be tested for, these propositions having more backing than those based on substrate similarity alone.

It is our hope that this ongoing work will provide the proof-of-concept for this particular method, and that it may be improved and applied to the other protein families of unknown function(s).

References

[1] A. Kreimeyer et al., "Identification of the Last Unknown Genes in the Fermentation Pathway of Lysine" J. of Bio. Chem., 10:7191-7197, 2007.

[2] R. D. Finn et al., "The Pfam protein families database" Nucleic Acids Res., 38:D211-222, 2010.

[3] A. Strehl, J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions" J. Mach. Learn. Res. 3:583-617, 2003.

[4] K. Katoh et al., "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform" Nucleic Acids Res., 30:3059-3066, 2002.

[5] K. Howe, A. Bateman, R. Durbin, "QuickTree: building huge Neighbour-Joining trees of protein sequences" Bioinformatics 18:1546-1547, 2002.

[6] D. P. Brown et al., "Automated Protein Subfamily Identification and Classification" PLoS Comput. Biol., 3:e160, 2007.

[7] D. Vallenet et al., "MicroScope: a platform for microbial genome annotation and comparative genomics" Database, 0:bap021, 2009.

[8] F. Boyer et al., "Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data" Bioinformatics, 21:4209-4215, 2005.

[9] U. von Luxburg, "A tutorial on spectral clustering" Statistics and Computing 17:395-416, 2007.

[10] R.C.M. Minardi, F. Artiguenave, "Homology modelling based protein functional residues prediction" ISMB Proceedings, 2009.

The Small, Slow and Specialized CRISPR and Anti-CRISPR of *Escherichia* and *Salmonella*

Marie TOUCHON[1,2] and Eduardo ROCHA[1,2]

[1] Institut Pasteur, Microbial Evolutionary Genomics, Département Génomes et Génétique, F-75015 Paris, France
[2] CNRS, URA2171, F-75015 Paris, France
mtouchon@pasteur.fr

Abstract Prokaryotes thrive in spite of the vast number and diversity of their viruses. This partly results from the evolution of mechanisms to inactivate or silence the action of exogenous DNA. Among these, CRISPR are unique in providing adaptive immunity against elements with high local resemblance to previously infecting genomes. Here, we analyze the CRISPR loci of 51 complete genomes of *Escherichia* and *Salmonella*. Our results match and extend previous analyses and by using phylogenetic analysis allowed us to propose evolutionary scenario for these systems. All CRISPR are in two pairs of loci, each pair showing a similar turnover rate, similar repeats and is associated with the same set of *cas* genes. Yet, we find evidence that CRISPR and associated *cas* genes have different evolutionary histories, with the latter being frequently changed or lost. One CRISPR pair seems specialized in plasmids often matching genes coding for the replication, conjugation and antirestriction machinery and, strikingly, the corresponding *cas* genes. We suggest that such anti-CRISPR can be used to counteract the invasion of mobile elements containing CRISPR. Unexpectedly, the number and turnover of spacers in these genomes seems incompatible with the expected dynamics of an immune system. Overall, these results suggest that enterobacterial CRISPR have complex roles providing a limited repertoire of defenses.

Keywords Comparative genomics, phages, plasmids, bacterial immunity, microbial evolution

1 Introduction

Prokaryotic viruses are the most abundant forms of life on Earth. Nevertheless, microbes routinely survive and thrive in remarkably phage-rich environments. This is because they have developed defense mechanisms that allow them to withstand viral predation and the constant exposure to exogenous nucleic acids. Recently, an adaptive microbial immune system, clustered regularly interspaced short palindromic repeats (CRISPR) has been identified that provides acquired immunity against any foreign DNA by targeting nucleic acid in a sequence-specific manner (see review [1]).

CRISPR have been identified in most genomes. CRISPR typically consist of short and highly conserved direct repeats regularly separated by stretches of variable sequences called spacers. 12 majors groups of CRISPR were defined based on sequence similarity of their repeats. CRISPR are often adjacent to *cas* (CRISPR-associated) genes. Cas proteins carry functional domains typical of nucleases, helicases, and polymerases, involved in the propagation and functioning of CRISPR. They were classified into 8 CRISPR/*cas* subtypes that often share gene order as well as content. CRISPR are typically preceded by an AT-rich non-coding sequence called leader. A new repeat-spacer (RS) unit is added to the CRISPR between the previous unit and the leader, which likely includes a binding site for the Cas proteins responsible for repeat duplication and/or spacer acquisition. The leader has also been proposed to act as a promoter for the transcription of the CRISPR array into a CRISPR transcript. A fully functional CRISPR/*cas* system is composed of the CRISPR, the Cas proteins and the leader. Previous studies have reported that many spacers of CRISPR derive from sub-sequences, named proto-spacers, of foreign mobile genetic elements (MGE). It has therefore been hypothesized that this system might be immunity-like systems. This role was first proved in 2007 in *S. thermophilus*: CRISPR-harboring strains became resistant to infection by phages after the acquisition of new spacers derived from the virus. It has also been shown that CRISPR/*cas* systems can limit plasmid conjugation in *S. epidermidis*, demonstrating a broader role for CRISPR in the prevention of HGT.

Here, we investigate the structure and

evolution of CRISPR in 51 complete genomes of *Escherichia* and *Salmonella*. These two genera include important pathogens and model bacteria. There is also substantial information for MGE in these genera. Most importantly, we try to understand the evolutionary history of CRISPR in a phylogenetic framework [2].

2 Results/Discussion

E. coli CRISPR have been identified before and the CRISPR1 has been well-described. Here, we report the characterization of the 3 other CRISPR. We confirmed several previous observations such that spacers are taken up randomly and non-directionally. When present in genomes, the CRISPR are always located at the same locations despite the multiple occurrences of *cas* genes degradation and *cas* horizontal transfer. This implies that the process of replenishing genomes with intact *cas* loci is frequent and that horizontally transferred *cas* genes are always inserted in the same locations, next to a given CRISPR. We propose that CRISPR might out-live the *cas* genes in the genome, thereby providing for an integration hotspot. This is most clearly demonstrated by the observation that sub-clades with different *cas* genes contain some similar spacers. We have shown that CRISPR1 and CRISPR2 on one hand and CRISPR3 and CRISPR4 on the other are functionally coupled: these pairs are co-localized in the genome, they have identical repeats, they are associated with similar CRISPR/*cas* genes subtypes, and tend to show correlated dynamics. CRISPR/*cas* subtypes might therefore be specialized and act in trans on all CRISPR with identical repeat sequences. Interestingly, the analysis of the spacers strongly suggests that CRISPR1 and CRIPSR2 target mostly phages, whereas CRISPR3 and CRISPR4 only target plasmids. Why and how CRISPR are specialized remains unknown but one could imagine different mechanisms aiming at responding to incoming DNA, dsDNA for phages and ssDNA for conjugative plasmids.

This study supports the idea that new spacers are acquired in a polarized fashion. This implies that spacers are chronological records reflecting previous encounters with MGE. However, the loss of one or more RS units has been observed. This suggests that CRISPR do not grow unchecked. One would assume that older spacers should be more frequently deleted because they have been inserted for a longer time. Surprisingly, some of them are highly persistent. This might indicate a critical unknown function in CRISPR/*cas* system activity. Our results also suggest that periods of *cas*-activity in the genome are associated with increase in CRISPR arrays and that the remaining periods are associated with the loss of spacers. Since these genomes contain relatively few spacers of which several are fairly ancient, and since CRISPR seem to change in a very irregular temporal pattern, the relevance of using these particular loci for typing and epidemiological studies is questionable.

CRISPR are consistently described as among the most rapidly evolving genomic loci. In our case, the CRISPR do not seem so hypervariable. No analyzed genome has more than 3 CRISPR. The CRISPR positions are strictly conserved and no locus has more than 34 RS units. In addition, strains that have diverged in the last thousand years have identical CRISPR showing a slow turnover of spacers relative to the generation time. This low dynamics of CRISPR in these genomes is puzzling since many MGE are known for these species. Despite, the outstanding opportunity provided by the availability of many sequenced enterophages, unlike in other clades, only 7% of these elements were matched by spacers. This means that these strains remain vulnerable to the vast majority of phages. This work seriously raises the question of CRISPR real efficiency in providing wide-range protection against enterophages.

Our results are consistent with previous reports on the high transmissibility of CRISPR and their association with MGE. Why CRISPR exist in MGE remains unknown but one could imagine their deleterious effects if they contain spacers matching the bacterial host. We are inclined to believe that residual CRISPR may confer selective advantages to their host cells and, in these cases, stabilizes the loci against degradation. This suggestion is strongly supported by the finding of a short CRISPR containing only spacers matching *cas* genes of its own subtype in all genomes devoid of the corresponding *cas* genes. We propose that CRISPR themselves can be used to prevent the invasion of MGE carrying functional CRISPR/*cas* system. This would be the first description of such an anti-CRISPR system. Our results provide an example of how evolutionary works using full closely related genome data might contribute to a comprehensive understanding of these intriguing elements.

References

[1] H. Deveau, JE. Garneau, S. Moineau. CRISPR/Cas System and Its Role in Phage-Bacteria Interactions. Annu Rev Microbiol, 2010.

[2] M. Touchon and EP. Rocha. The Small, Slow and Specialized CRISPR and anti-CRISPR of Escherichia and Salmonella. PLoS One,15;5(6):e11126, 2010.

Differentiation of allelic frequencies analysis identifies short genomic regions with signatures of artificial selection between canine breeds

Amaury VAYSSE[1], Abhirami RATNAKUMAR[2], Thomas DERRIEN[1], Kerstin LINDBLAD-TOH[2,3], Catherine ANDRE[1], Matthew T. WEBSTER[2*] and Christophe HITTE[1*]

[1] IGDR, UMR6061 CNRS, 2 av Pr. L.Bernard, 35043, Rennes, France
{amaury.vaysse, thomas.derrien, catherine.andre, christophe.hitte}@univ-rennes1.fr
[2] Department of Medical Biochemistry and Microbiology, Uppsala University, Box 582, SE-751232 Uppsala, Sweden
{abhirami.Ratnakumar, matthew.webster}@imbim.uu.se
[3] Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA
kersli@broadinstitute.org

*equal contribution

Abstract *The canine species has been domesticated for 15,000 years and is today composed of ~350 distinct breeds that result from intense artificial selection and breeding practices by human during the last centuries. However, patterns of genetic variation that indicate such recent selection events and the underlying functional mutations remain unknown. Here, we describe a method that detects differentiation of allelic frequencies and apply it to a genome-wide scan for signatures of artificial selection using 510 dogs from 31 distinct breeds genotyped with more than 170,000 SNPs. The method based on the variance of allelic frequency using Wright's Fixation index (F_{ST}) statistic, pinpoints 2500 short genomic regions (mean 200 kb) that possess a robust genetic signature. Interestingly, we identify several known genes such as IGF1 involved in breed size differentiation, or HAS2 which is associated to skin-wrinkle Shar-pei phenotype and 577 new candidate genes. In a given breed, the signatures span less than 1% of the genome and localize ~80 candidate genes that will help to understand on which genetic and functional basis size, coat-type, morphology or behavior have been selected. Our results not only identify individual genes and short polymorphism targets but also reveals instances of functional categories showing signs of artificial selection that are distinct from natural selection. We have established a highly resolutive map of recent signatures of selection of the canine genome that pinpoint new functional candidates to account for most differentiated phenotypes in purebred dogs.*

Keywords Evolution, Selection signatures, Dog, Fst, SNP

L'analyse de la différentiation allélique identifie de courtes régions génomiques avec signatures de sélection artificielle entre races canines

Résumé *L'espèce canine domestiquée il y a près de 15000 ans, se compose aujourd'hui de plus de 350 races qui résultent de la sélection artificielle et des pratiques d'élevage appliquées par l'homme essentiellement ces deniers siècles. Cependant, les patrons de variation génétique qui reflètent ces événements de sélection récents, et donc les gènes sous-jacents et leurs mutations, sont encore largement inconnus. Dans ce travail, nous décrivons une méthode d'identification de signatures de sélection artificielle basée sur la différentiation allélique entre races. Nous avons appliqué cette méthode sur le génome canin en analysant 510 chiens de 31 races distinctes génotypés avec plus de 170000 SNP. Cette analyse basée sur le calcul de la variance des fréquences alléliques par l'index de Fixation de Wright (Fst) localise plus de 2500 courtes régions génomiques (~200kb) qui possèdent une signature de sélection robuste. Nous identifions plusieurs gènes tel que IGF1 connu pour son implication dans la différentiation de la taille des races ou HAS2 qui est associé au phénotype peau plissé du Shar-pei ainsi que 577 robustes nouveaux candidats. Pour une race donnée, l'ensemble des signatures cible moins de 1% du génome et identifie ~80 gènes qui permettront d'aider à déterminer les bases génétiques et fonctionnelles de la sélection de phénotypes tels que la taille, le pelage, la morphologie ou de comportement. Nos résultats identifient non seulement des polymorphismes et des gènes uniques mais révèlent également des catégories fonctionnelles candidates qui signent la sélection artificielle bien distinctes des classes fonctionnelles détectées dans les événements de sélection naturelle. Nous établissons une cartographie très résolutive des signatures de sélection artificielle du génome canin, qui identifie de nouveaux candidats fonctionnels pouvant contribuer aux phénotypes les plus différentiés entre races et groupe de races chez le chien.*

Mots-clés Evolution, Signatures de sélection, Chien, Fst, SNP

1 Introduction

Dans leur infinie variété, toutes les chiens descendent du loup gris (*Canis lupus*) domestiqué par l'homme il y a environ 15000 ans [1, 2, 3], et se seraient répandus dans toute l'Asie et l'Europe, avant d'accompagner l'homme dans le nouveau monde. Le processus de domestication correspond à un premier goulet d'étranglement de l'histoire évolutive du chien et a eu pour conséquence d'intensifier le processus de dérive génétique à partir d'un pool relativement restreint d'allèles. Un second goulet d'étranglement coïncide avec la création de plus de 350 races canines par l'homme au cours des deux derniers siècles. Les pratiques de sélection intensive ont été mises en œuvre par des générations d'éleveurs qui ont croisé et sélectionné des animaux avec pour finalité de disposer de races possédant des morphologies particulière, des aptitudes d'intérêt dédiées à la garde ou à la chasse par exemple. [4].

En conséquence, cette spectaculaire diversité de l'espèce canine engendrée en quelques siècles suggère une composante génétique forte de la variabilité phénotypique. Ainsi, l'homogénéité du phénotype d'une race reflète une forte homogénéité génétique alors que la diversité entre races suggère la présence de signatures génétiques laissées par la sélection artificielle. Contrairement à la sélection naturelle qui agit au cours de l'évolution pendant des millions d'années, la sélection artificielle telle qu'elle est pratiquée chez les espèces domestiquées est un processus rapide qui façonne l'architecture génétique des races créées en fixant des patrons de polymorphisme. L'identification des événements de la sélection artificielle repose principalement sur deux méthodes ; 1- *les haplotypes étendus* ; un allèle sélectionné augmente sa fréquence si rapidement que son association avec les polymorphismes voisins n'est pas réarrangée par la recombinaison ; 2- *les allèles fortement différenciés ;* un allèle sélectionné dans une population cause une plus grande différence de fréquence entre populations que pour des allèles sous évolution neutre.

La structuration de l'espèce en races, la disponibilité de la séquence complète de son génome, de 2,5 millions de SNPs [5], et d'outils génomiques tels que les puces d'expression et de SNPs font du chien un modèle de choix pour la recherche de signature de sélection artificielle. Paradoxalement, des études exhaustives et résolutives n'ont pas été réalisées pour identifier les signatures génétiques, les patrons de polymorphisme et les variants fonctionnels qui contribuent à la différentiation des races. La plupart des études a été guidé par l'étude d'un phénotype d'intérêt et a porté sur l'analyse de gènes isolés. Seuls cinq locus impliquant cinq variants fonctionnels ont été associés à la variabilité entre races pour les phénotypes de taille, de texture et de longueur du pelage et de morphologie du squelette [6, 7]. Une étude plus récente [8] a localisé ~150 vaste loci (1 Mb) candidats de sélection artificielle qui cependant, ne permettent pas de cibler avec précision quels variants fonctionnels et quels gènes sont réellement impliqués.

Dans cette étude, nous avons analysé l'ensemble du génome canin par le génotypage de 170 000 SNPs (1 SNP/15kb) sur 510 chiens appartenant à 31 races distinctes. Chaque race comprend 16 individus en moyenne avec un minimum de 10 chiens et un maximum de 25 et 26 pour 6 races. Nous avons recherché les régions génomiques qui possèdent des patrons de polymorphismes fortement différenciés entre races et ainsi identifient des signatures de sélection liées à la création des races canines. Pour chaque SNP nous avons calculé l'indice de fixation de Wright 'F_{st}' pour chaque paire de race et dérivé une métrique 'di' qui mesure la valeur de la variance des fréquences alléliques de chaque SNP qui différencie chaque race de toutes les autres.

Plus de 2500 régions d'une taille moyenne de 200 kb, contenant en moyenne deux gènes ont été identifiées avec une signature robuste de sélection artificielle. Nous avons détecté plus de 1800 nouveaux loci qui identifient une signature de sélection spécifique de races et 676 locus qui identifient une signature partagée entre groupe de races. Un total de 577 régions localisent un gène unique et ciblent ainsi un variant fonctionnel candidat unique pour être associé à un phénotype fortement différentié entre races. Nous avons déterminé que pour une race donnée, l'ensemble des signatures spécifiques occupe moins de 0.5% du génome et comprend environ 60 locus qui ciblent avec précision des candidats fonctionnels contribuant à la différentiation des races.

2 Résultats

Nous avons génotypé plus de 170 000 SNPs autosomiques avec la technologie des puces à ADN Illumina Infinium CanineSNP170 sur une cohorte de 510 chiens non apparentés de 31 races distinctes phénotypiquement. Les SNPs sont uniformément répartis sur l'ensemble du génome, avec une densité moyenne de 1 SNP tous les 15 kb. La qualité du génotypage a été évaluée par un taux de concordance du 'call rate' >99% entre réplicats. La totalité

des ~87 millions de génotypes a été incluse dans l'analyse. Les SNP non polymorphes ont été identifiés par le programme Haploview [9] sur l'ensemble des 510 individus de la cohorte et ont été éliminés des analyses ultérieures en raison de leur absence d'informativité. L'ensemble de ces filtres a garanti un jeu de données de génotype polymorphes de haute qualité.

2.1 Stratégie et validation de l'identification de signatures de sélection artificielle dans le génome canin

La stratégie utilisée dans cette étude se base sur une approche de génétique des populations qui mesure le niveau de différentiation allélique entre populations afin de détecter les patrons de polymorphisme qui signent une population. Pour chaque SNP (n=170000), nous avons calculé l'indice F_{st} de fixation = $(\prod_{between} - \prod_{within})/\prod_{between}$ pour chacune des 465 combinaisons de paires de races. Pour comparer chaque valeur de F_{st}, nous avons déterminé une valeur statistique dérivée 'di' qui est une fonction des valeurs de Fst par paire entre une race 'i' et l'ensemble des autres races telle que : $di=\sum_j(F_{st}^{ij}-E(F_{st}^{ij}))/sd(F_{st}^{ij})$ où $E(F_{st}^{ij})$ et $sd(F_{st}^{ij})$ représentent respectivement la valeur attendue et l'écart-type du F_{st} entre les races i et j calculée à partir de la totalité des SNPs. Afin de limiter la variation aléatoire du polymorphisme allélique d'un seul SNP donc d'une valeur individuelle de di, nous avons moyenné les valeurs de di par fenêtre de 150 kb glissantes avec un pas de 25 kb. Une fenêtre de 150 kb permet de considérer en moyenne 10 SNPs (soit 10 valeurs de di), toutes les fenêtres contenant moins de 5 SNPs ont été écartées de l'analyse. A partir des valeurs de di, nous avons définis les fenêtres sous sélection comme étant les valeurs 'outliers' de la distribution des di qui dépassent le 95ème percentile de cette distribution. A partir des 87766 fenêtres glissantes analysées sur l'ensemble du génome, 9234 fenêtres ont été détectées comme 'outliers' de la distribution et ont pu être regroupées lorsqu'elles étaient chevauchantes en 2503 régions génomiques. Plusieurs contrôles suggèrent la validité des régions identifiées sous sélection. Une précédente étude [8] a décrit 155 vastes loci génomiques de 1 Mb contenant 1630 gènes (~11 gènes par locus) candidats à la sélection artificielle canine. Ces locus sont en grande majorité (79%) identifiés par notre étude et 60% sont réduits à une région génomique inférieure à 400 kb qui cible 2.8 gènes en moyenne en comparaison des 11 gènes contenu par locus localisés par Akey et al. Par ailleurs, cinq locus de différentiation de races canines ont été préalablement caractérisés et concernent le phénotype 'petite taille' associé au gène IGF1 [10], le phénotype de morphologie du squelette qui confère des membres très courts et achondrodisplasiques à certaines races (teckel) pour lequel l'implication d'un rétrogène de FGF4 a été démontré [7], deux gènes (RSPO2, KRT71) impliqués dans la différentiation du pelage entre races [6] et le phénotype de 'peau plissée' du Shar-pei pour lequel le gène HAS2 a été associé [8]. Les régions localisées dans notre étude identifient avec précision les cinq locus qui contiennent les cinq gènes connus pour leur implication dans la différentiation des races.

2.2 Caractérisation des régions génomiques avec signatures de sélection

Les 2503 régions candidates canine de sélection artificielle identifiées dans cette étude se répartissent sur l'ensemble du génome et ont une taille de 200 kb en moyenne. La taille cumulée des régions s'élève à 479 Mb représentant 19% du génome canin. Une densité plus élevée en gènes codant pour des protéines a été détectée dans les régions candidates (wilcoxon test , p=0.003) suggérant que la sélection agit sur les éléments fonctionnels. Afin d'évaluer leur contenu en éléments fortement conservés au cours de l'évolution, nous avons assigné, pour chacune des 2503 régions un score de conservation de séquence issu de l'alignement multiple de séquences codant pour des protéines entre homme, souris, rat et chien calculé par le programme phastcons [11]. Nous avons relocalisé les 2503 régions aléatoirement sur le génome canin, en respectant leurs tailles originales, et recalculé les scores de conservations des séquences codantes. La comparaison des 2 distributions de scores ne montre pas de différence significative (t-test p= 0.4156) entre les régions sous sélection artificielle et les régions prises aléatoirement dans le génome canin suggérant que la sélection artificielle opérée sur les races de chien est un processus qui ne se restreint pas aux séquences fortement conservées au cours de l'évolution telles que les régions codantes.

A partir des 19014 gènes canins codant pour des protéines annotés par le serveur Ensembl, l'ensemble des régions sous sélection contient 3458 gènes. Un total de 923 gènes sont localisés dans une signature identifiée dans deux races au moins et peuvent permettre d'analyser sur quelles bases génétiques et fonctionnelles les groupes de races proches, tels que les retrievers, les petits terriers, les molosses, ont été crées. Les 2535 gènes restants sont localisés dans des signatures strictement spécifiques de races et constituent de bons candidats qui contribuent à la spécificité génétique et fonctionnelle de la race et/ou participer à la restriction de la variabilité du phénotype au sein d'une race.

Plus de la moitié des régions localisées dans notre étude (n=1343) contiennent des gènes codant pour

des protéines alors que plus de 1100 régions sont dépourvues de gènes annotés. De nombreuses signatures chez l'homme ont été localisées dans les régions intergéniques [12], suggérant que les variants sélectionnés peuvent correspondre à des éléments régulateurs tels que les régions promotrices, des sites de liaisons aux facteurs de transcription ou des ARN non codants pour des protéines. Par ailleurs, l'absence de gènes connus associées aux signatures de sélections identifie des régions génomiques pour lesquelles un effort de réannotation du génome canin doit être considéré.

2.3 Signatures de la sélection spécifique de races et partagées entre races

Parmi les 2503 régions candidates possédant une signature de la sélection, 73% (n =1827) sont détectées dans une seule race. Dans l'exemple d'une région du chromosome 13, l'étude détecte spécifiquement dans la race shar-pei un locus de 187 kb (Fig. 1). Ce locus contient un gène unique HAS2 qui a été récemment identifié [8] comme associé au phénotype 'peau plissée' présent dans la race Sharpei et absent dans 30 autres races de la cohorte. Dans cet exemple le gène HAS2 fait partie des meilleurs candidats qui peuvent être identifié par une approche qui intègre les données de signatures spécifiques de la race Shar-pei avec des données de type gène candidat pour lesquels, une corrélation entre le phénotype 'peau plissée' et une fonction liée à la physiologie et au métabolisme de la peau est recherchée.

Au delà de la localisation de patrons de polymorphisme isolés et de gènes individuels candidats, notre approche permet d'identifier l'ensemble des signatures de sélection spécifique d'une race. Nous avons déterminé la combinatoire des locus qui contribue pour chaque race à leur différentiation de l'ensemble des autres races. Pour chaque race, la combinatoire de locus spécifiques implique en moyenne 0.4% du génome (~10 Mb) et cible 81 gènes codant pour des protéines. Ces gènes constituent des candidats fonctionnels ciblés par la sélection artificielle et pouvant contribuer aux phénotypes les plus différentiés entre races et groupe de races chez le chien.

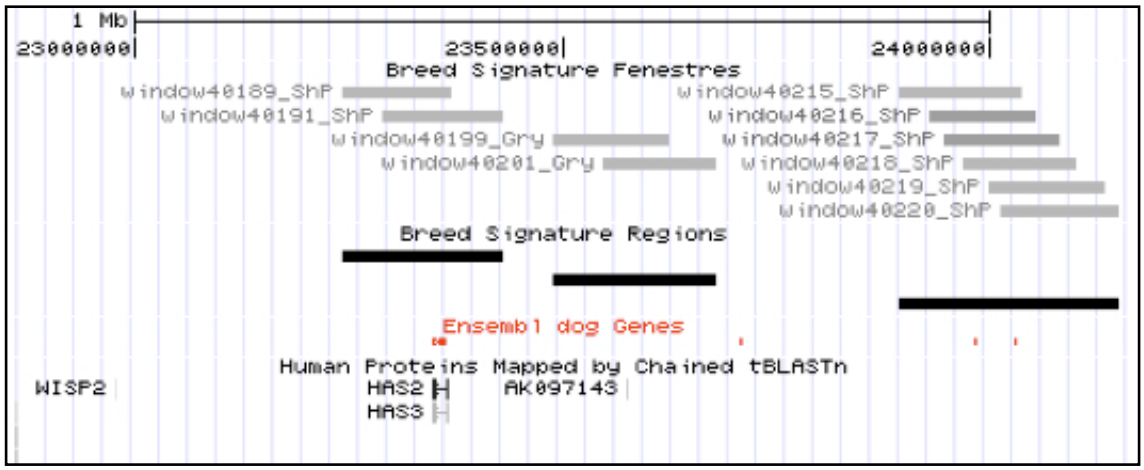


Fig. 1. Représentation de locus génomique de sélection artificielle : Exemple du chromosome 13

La méthode repose sur l'identification de fenêtres glissantes de 150 kb (barres horizontales grises) par races canine pour lesquelles une valeur seuil de statistique dérivée du F_{st} est retenue. La réunion des fenêtres glissantes chevauchantes en région est illustrée par les barres noires horizontales. La représentation des fenêtres et des régions dans le contexte du serveur UCSC permet de visualiser le contenu en gènes canin et humain pour chaque régions candidates sous sélection artificielle.

Environ 27% (n=676) des signatures de sélection sont observées entre au moins deux races. De telles signatures suggèrent que l'action d'un gène ou d'un élément fonctionnel réunit des races en groupe de race et en classe de phénotypes convergeant. Par exemple une signature détectée dans notre étude contient le gène IGF1 qui est associé à la taille et gouverne le phénotype 'taille miniature' [10]. Nous retrouvons la signature dans l'ensemble des races du groupe 'toy' ainsi que d'autres races du groupe 'géant' correspondant à des phénotypes très différenciés. Pour le phénotype correspondant au pelage de type 'fournis' gouverné par le gène RSPO2 [6], une signature est identifiée dans les 4 races affichant ce trait et contenant effectivement RSPO2. Ainsi une signature partagée entre races permet de

dresser l'inventaire des gènes qui peuvent être associé à des phénotypes fixés entre groupe de race et fortement différenciés d'autres groupes de races comme la morphologie du crâne, la musculature ou associés à des aptitudes fixées dans un groupe de races telles que la chasse chez les retrievers ou la course chez les lévriers.

2.4 Analyse fonctionnelles des gènes de sélection artificielle

Les 2503 régions détectées avec une signature de sélection contiennent plus de 3400 gènes annotés et prédits pour coder des protéines. Afin de structurer les relations entre gènes et d'identifier les processus biologiques et les fonctions moléculaires impliquées dans les signatures observées, nous avons analysé leur annotation fonctionnelle par les termes GO (Gene Ontology) des gènes humains orthologues [13]. Nous avons utilisé les gènes (n=294) ayant une relation d'orthologie de type 1:1 entre l'homme et le chien, des régions qui détectent un gène unique. Pas de biais significatif d'échantillonnage lié a l'exclusion de famille de gènes (Test chi-deux, p=0.87) ou d'orthologues 1:1 (Test chi-deux, p=0.37) n'a pu être détecté. L'analyse GO a permit d'analyser le plus précisément possible la fonction ciblée par la sélection artificielle et de la discriminer du bruit apporté par des gènes non directement ciblés par la sélection dans le cas des loci qui détectent plusieurs gènes. Les catégories fonctionnelles retrouvées significativement enrichies ($p<0.001$) par l'analyse GO appartiennent aux classes de la prolifération cellulaire, du développement des organes, de la régulation des processus physiologiques et cellulaires pour les termes de processus biologique et à la classe fonctionnelle des protéines qui se lient aux facteurs de transcription pour les termes de fonction moléculaire. Certaines fonctions moléculaires conservées impliquants les protéines du cytosquelette et la motilité et la structure de la cellule ont été identifiées sous faible sélection négative chez l'homme [14]. Cependant les catégories fonctionnelles identifiées dans ce travail sont en grande partie distinctes de celles retrouvées classiquement liées à sélection naturelle par sélection positive, telles que les fonctions impliquées dans l'immunité, la défense de l'organisme, les processus biologiques de réponse aux stimulus [15]. Par exemple, la catégorie enrichie en gènes se liant aux facteurs de transcriptions suggère que la régulation de l'expression des gènes peut être significativement impliquées dans les phénomènes de sélection artificielle au même titre que les gènes impliqués dans la structure des protéines. Parmi les facteurs de transcription, des gènes homéobox ont été identifiés dont HOX11L2 qui code pour un facteur de transcription nucléaire de liaison à l'ADN et TBX5 qui code un facteur de transcription impliqué dans la régulation des processus du développement. Le gène MITF identifié sous sélection régule la différentiation et le développement des mélanocytes.

2.5 Discussion

Nous décrivons ici l'identification de régions de différentiation allélique entre races canines par une approche statistique basée sur le calcul de la variance des fréquences alléliques par l'index Fst. Le choix de dériver un index 'di' qui somme des valeurs centrées réduites nous permet de générer des valeurs de même dispersion par race et ainsi de pouvoir les comparer. L'utilisation d'une fenêtre de 8 à10 SNPs a pour but de limiter la détection de faux-positifs par rapport à une approche qui considérerait une valeur unique. Il serait utile cependant de tester la corrélation entre les résultats obtenus avec les valeurs moyennées par fenêtre par rapport aux valeurs individuelles. Le choix d'un seuil correspondant au 95 ème percentile des distribution de 'di' est empirique. La possibilité de réaliser des tests de permutations en faisant varier les identifiants de races permettra de disposer d'une distribution théorique des valeurs de Fst, à partir de laquelle nous testerons la significativité de la valeur observée.

Par opposition aux études d'association de type cas-contrôles qui nécessitent la connaissance du phénotype à tester pour identifier les marqueurs associés [16], notre approche permet de déterminer les régions génomiques qui possèdent une forte différence de patrons de polymorphisme sans connaissance *a priori* du phénotype. La connaissance précise du phénotype pour chaque race serait cependant très informative pour tester une corrélation phénotype-génotype. L'inclusion de paramètres quantitatifs et qualitatifs d'ordre morphologiques, physiologiques ou comportementaux pourrait permettre de rechercher une corrélation statistique entre un ou plusieurs paramètres et les gènes et éléments fonctionnels identifiés dans les régions.

La possibilité de coupler une étude d'association à partir des données obtenues par l'approche F_{st} est une idée qui peut sembler séduisante mais qui souffre de la nécessité de disséquer la variation du phénotype et d'attribuer un phénotype aux populations à tester. Une approche préalable de classification par clustering hiérarchique des races identifiées par une signature de sélection peut permettre de distinguer les phénotypes auxquelles les statuts cas et contrôles seront attribués.

L'analyse des haplotypes étendus (EHH) [17] par laquelle on recherche si un allèle sélectionné va augmenter sa fréquence assez rapidement pour que

son association avec les polymorphismes voisins ne soit pas altérée par la recombinaison, devrait permettre de localiser de manière indépendante les loci soumis à un événement de sélection très récent. Nous anticipons que l'intégration des résultats de notre étude avec des données d'analyse d'EHH facilitera et renforcera la mise en évidence de locus candidats de sélection artificielle. La meilleure définition des régions permettra d'extraire de manière plus précise les gènes ciblés et ainsi de mener des analyses fonctionnelles par termes GO sur des effectifs plus importants.

Le développement de cette approche a permis de valoriser les données de polymorphisme du génome canin, de déterminer de nouveaux loci génomiques et de nouveaux candidats fonctionnels pouvant contribuer aux phénotypes les plus différentiés entre races et groupe de races chez le chien. A terme, nous anticipons que la comparaison des mécanismes et des fonctions impliquées dans les événements de sélection pourront faciliter la compréhension des processus de sélection naturelle et artificielle chez le chien et que cette approche pourra être appliquée aux autres espèces domestiquées.

Remerciements

Nous remercions la plate-forme de bio-informatique Genouest pour leur aide technique et leur assistance. Ce travail est en partie financé par la commission Européenne (LUPA - GA-201370). Nous remercions le CNRS pour le support financier apporté à AV, TD, CA et CH, et l'université de Uppsala pour AR, KLB et MW.

Références

[1] C. Vila, P. Savolainen, JE. Maldonado, IR. Amorim, JE. Rice, RL. Honeycutt, KA. Crandall, J. Lundeberg and RK Wayne, Multiple and ancient origins of the domestic dog. Science, 276(5319):1687-1689, 1997.

[2] RK. Wayne and EA. Ostrander, Origin, genetic diversity, and genome structure of the domestic dog. Bioessays, 21(3):247-257, 1999.

[3] BM. Vonhold, JP. Pollinger, KE. Lohmueller, E. Han, HG. Parker, P. Quignon, JD. Degenhardt, AR. Boyko, DA. Earl, A. Auton, A. Reynolds, K. Bryc, A. Brisbin, JC. Knowles, DS. Mosher, TC. Spady, A. Elkahloun, E. Geffen, M. Pilot, W. Jedrzejewski, C. Greco, E. Randi, D. Bannasch, A. Wilton, J. Shearman, M. Musiani, M. Cargill, PG. Jones, Z. Qian, W. Huang, ZL. Ding, YP. Zhang, CD. Bustamante, EA. Ostrander, J. Novembre and RK. Wayne, Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature, Apr 8;464(7290):898-902, 2010.

[4] F. Galibert F and C. Andre, The dog genome. Genome Dyn, 2:46-59, 2006.

[5] K. Lindblad-Toh, C.M. Wade, T.S. Mikkelsen, E.K. Karlsson, D.B. Jaffe, M. Kamal, M. Clamp, J.L. Chang, E.J. Kulbokas, M.C. Zody et al., Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature, 438:803-819, 2005.

[6] E. Cadieu, MW. Neff, P. Quignon, K. Walsh, K. Chase, HG. Parker, BM. Vonholdt, A. Rhue, A. Boyko, A. Byers, A. Wong, DS. Mosher, AG. Elkahloun, TC. Spady, C. André, KG. Lark, M. Cargill, CD. Bustamante, RK. Wayne and EA. Ostrander, Coat variation in the domestic dog is governed by variants in three genes. Science, 326(5949):150-3, 2009.

[7] HG. Parker, BM. VonHoldt, P. Quignon, EH. Margulies, S. Shao, DS. Mosher, TC. Spady, A. Elkahloun, M. Cargill, PG. Jones, CL. Maslen, GM. Acland, NB. Sutter, K. Kuroki, CD. Bustamante, RK. Wayne and EA.Ostrander. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science, 325(5943):995-8, 2009.

[8] JM. Akey, AL. Ruhe, DT. Akey, AK. Wong, CF. Connelly, J. Madeoy, TJ. Nicholas and MW Neff, Tracking footprints of artificial selection in the dog genome. Proc Natl Acad Sci U S A, 107(3):1160-5, 2010.

[9] JC. Barrett, B. Fry, J. Maller and MJ. Daly, Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics; 15;21(2):263-5, 2005.

[10] NB. Sutter, CD. Bustamante, K. Chase, MM. Gray, K. Zhao , L. Zhu, B. Padhukasahasram, E. Karlins, S. Davis, PG. Jones, P. Quignon, GS. Johnson, HG. Parker, N. Fretwell, DS. Mosher, DF. Lawler, E. Satyaraj, M. Nordborg, KG. Lark, RK. Wayne RK and EA. Ostrander, A single IGF1 allele is a major determinanat of small ize in dogs. Science, 316(5821):112-5, 2007.

[11] A. Siepel, G. Bejerano, JS. Pedersen, AS. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, LW. Hillier S. Richards, GM. Weinstock, RK. Wilson, RA. Gibbs, WJ. Kent, W. Miller and D. Haussler. Evolutionarily conserved elements in vertebrate insect worm and yeast genomes, Genome Res., 8:1034-50, 2005.

[12] SR. Grossman, I. Shylakhter, EK. Karlsson, EH. Byrne, S. Morales, G. Frieden, E. Hostetter, E. Angelino, M. Garber, O. Zuk, ES. Lander, SF. Schaffner and PC. Sabeti, A composite of multiple signals distinguishes causal variants in regions of positive selection. Science, 327(5967):883-6, 2010.

[13] B. Zhang, D. Schmoyer, S. Kirov and L. Snoddy, GOTree (GOTM) machine : a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. BMC Bioinformatics, 5:16, 2004.

[14] CD Bustamante, A Fledel-Alon, S Williamson, R. Nielsen, M. Todd-Hubish, S. Glanowski, DM. Tanenbaurm, TJ White, JJ. Sninsky, RD. Hernandez, D Civello, MD Adams, M. Cargill and AG Clark, Natural selection on protein-coding genes in the human genome. Nature, 437(706):1153-7, 2005.

[15] T. Derrien, J. Thézé, A. Vaysse, C. André, EA. Ostrander, F. Galibert and C. Hitte, Revisiting the missing protein-coding gene catalog of the domestic dog. BMC Genomics, 10:62, 2009.

[16] NA. Rosenberg, L. Huang, EM. Jewett, ZA. Szpiech, I. Jankovic and M. Boehnke, Genome-wide association studies in diverse populations. Nat Rev Genet., 5:356-66, 2010.

[17] PC. Sabeti, DE. Reich, JM Higgins, HZ. Levine, DJ. Richter, SF. Schaffner, SB. Gabriel, JV. Platko, NJ. Patterson, GJ. McDonald, HC. Ackerman, SJ. Campbell,

D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward and ES. Lander, Detecting recent positive selection in the human genome from haplotype structure. Nature, 6909:832-7, 2002.

ppALIGN: posterior probabilities for score-based alignments

Stefan WOLFSHEIMER[1], Alexander K HARTMANN[2] and Gregory NUEL[1]

[1] MAP5 - Mathématiques Appliquées à Paris Descartes, UMR 8145 CNRS, 45 rue des Saints Pères, 75270 Paris Cedex 06, France
gregory.nuel,stefan.wolfsheimer@parisdescartes.fr
[2] Carl von Ossietzky Universität Oldenburg Institut für Physik, D-26111 Oldenburg, Germany
alexander.hartman@uni-oldenburg.de

Keywords Sequence alignment, posterior probabilities, hidden Markov model

1 Introduction

Many search tools for large molecular database, like the BLAST family [3], rely on score-based alignments due to the existence of fast search heuristics. For a given query, a list of alignments is reported in descending order of the significance.

Since score based aligners produce unique alignments that maximize the score, they lack in a statistical analysis of the accuracy of the produced alignment. So as to address this problem, various probabilistic alignment methods, such as pair hidden Markov Models (pair-HMMs), have been developed in the last decade [8]. They provide a statistical description of the set of all alignments for a given pair of sequences including alternative meaningful alignments that may be hidden behind the optimum.

A classical model, termed "finite-temperature alignment", introduced in 1995, [4,5,9] gives to alignments the weight of an exponential (or Boltzmann) distribution. It can readily be applied to any classical scoring function with one additional parameter, the temperature T. For the canonical value $T = 1$, and using an appropriate scaled score matrix, it approximates more complex probabilistic models quite well [4,12].

Lunter et. al. [11] illustrated the usefulness of probabilistic alignment in detecting regions of low confidence. Especially close to gaps (i. e. insertions or deletions) often many competitive alignments decrease the accuracy of the maximum score alignment. These biases have been identified as "gap wander" [6], "gap attraction" and "gap annihilation" [10]. Gap wander describes the effect that an inferred gap position is shifted by a few pairs with respect to the "true alignment". Gap attraction occurs when two closely distant gaps merge into a single gap in the inferred alignment and the third effect is a cancellation of an insertion and a deletion.

In this presentation, we show that the software `ppALIGN` [12] can be useful in the posterior analysis of score based alignments. The software can process either a single alignment or the entire output of BLAST.

2 Methods and Results

Pairwise sequence alignment is a method to arrange letters from a pair of sequence $a_1^\ell = a_1 \dots a_\ell \in \Sigma^\ell$ and $b_1^m = b_1 \dots b_m \in \Sigma^m$ [8] in a way that the specific order of the sequences is preserved. More specifically, alignment algorithms aim at identifying regions of high similarity, because those regions are most likely related by evolution. Score based methods determine the optimal alignment $\hat{\pi}$ by maximizing an objective function s, $\hat{\pi} = \operatorname{argmax}_\pi s(\pi; a_1^\ell, b_1^m)$. The Needleman-Wunsch or the Smith-Waterman algorithm [1,2] are commonly used for global or local alignment respectively.

Probabilistic alignment methods go beyond the optimum and consider the set of possible alignments weighted with the so called posterior distribution

$$\mathbb{P}\left(\Pi = \pi \,\middle|\, a_1^\ell, b_1^m\right). \tag{1}$$

In cases where the optimal alignment agrees undoubtedly with the true (unknown) alignment, virtually all weight is put on the optimal alignment. When less similar sequences are compared to each other there might be regions of low confidence where letters might be aligned incorrectly or gaps are misplaced. The posterior distribution Eq. 1 is appropriate to quantify the degree of confidence for a given alignment.

`ppALIGN` uses pair-HMM techniques (or alternatively the finite temperature approach) to marginalize the posterior distribution of Eq. 1 and determine *column-wise posterior probabilities* [8]. The user may choose either global or local alignment models.

Let us assume that the optimal alignment relates position a_i in the first sequence to the position b_j in the second sequence. The confidence that this pair is

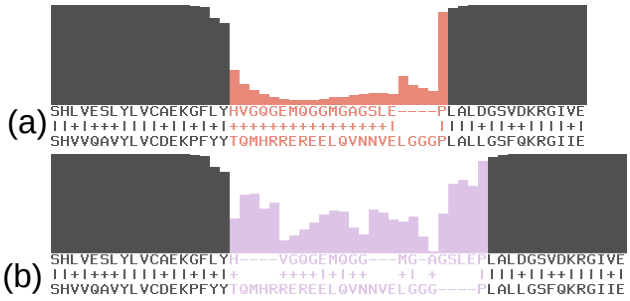


Fig. 1. Posterior probabilities of global alignments of simulated sequence pairs The posterior probabilities are based on `blosum62` with gap open penalty of $d = 11$ and extension penalty of $e = 1$. (a) The optimal alignment (b) Sampled alignment. Note that the aligned pair of Ps at the end of the low confidence region does not appear in the correct alignment, even though pairing yields a larger score.

aligned correctly can be assessed by the marginal posterior probability

$$P^{\mathsf{P}}_{i,j} = \mathbb{P}\left(\pi : a_i \text{ and } b_j \text{ paired } \middle| a_1^{\ell}, b_1^{m}\right) \qquad (2)$$

Similarly, we may also assign probabilities to gapped positions.

Example. In Fig. 1 we illustrate a typical example of the output of `ppALIGN`, (a) the optimal alignment and (b) an alignment sampled from the posterior distribution. The posterior probabilities are indicated by vertical bars. So as to evaluate the ability of `ppALIGN` to identify uncertain regions of global alignments, we have performed computer simulations which generated families of random sequences using the ROSE model of random sequence evolution [7]. The regions where the alignments in Fig. 1 agree with the "correct" (typically unkown) alignment are shown in a darker greyscale. Obviously, the regions of low confidence (with small posterior probabilities) also agree with the regions where the optimal alignment fails to predict the true alignment (see also [12] for a systematic study of the Receiver Operating Characteristic).

3 Conclusions

The package `ppALIGN` [12] (including stand-alone command-line programs and a C++ library) provides efficient algorithms that compute the posterior probabilities for score-based alignment. One stand-alone program, `ppALIGN`, allows the user to provide a single alignment and the set of parameters. The other one, `ppBLAST`, directly uses the structured output of BLAST (XML-format). It computes the posterior probabilities for each alignment. Both programs allow for a structured output in the XML format, plain text, and a more visual HTML page. The flexible library can be extended towards new decoding algorithms and other ways of marginalization of the posterior distribution.

References

[1] Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *J. Mol. Biol.* 1970, **48**:443–453.

[2] Smith TF, Waterman MS: **Identification of Common Molecular Subsequences**. *J. Mol. Biol.* 1981, **147**:195–197.

[3] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool**. *J. Mol. Biol.* 1990, **215**:403–410.

[4] Miyazawa S: **A reliable sequence alignment method based on probabilities of residue correspondences**. *Protein Eng.* 1995, **8**(10):999–1009.

[5] Zhang M, Marr T: **Alignment of Molecular Sequences Seen as Random Path Analysis**. *J. Theor.Biol.* 1995, **174**:119–129.

[6] Holmes I, Durbin R: **Dynamic Programming Alignment Accuracy**. *Journal of Computational Biology* 1998, **5**(3):493–

[7] Stoye J, Evers D, Meyer F: **Rose: generating sequence families**. *Bioinformatics* 1998, **14**:157–163

[8] Durbin R, Eddy S, Krogh A, Mitchison G: *Biological Sequence Analysis*. Cambridge University Press 1998.

[9] Kschischo M, Lässig M: **Finite-temperature sequence alignment**. In *Pacific Symposium on Biocomputing 5* 2000.

[10] Lunter G: **Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes**. *Bioinformatics* 2007, **23**(13):i289–296.

[11] Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J: **Uncertainty in homology inferences: Assessing and improving genomic sequence alignment**. *Genome Research* 2008, **18**(2):298–309.

[12] Wolfsheimer S, Hartmann AK, Nuel G: **ppALIGN: posterior probabilities for score-based alignments**. *Under review* 2010. Software available at `ppalign.sourceforge.net` Server: `www.mi.parisdescartes.fr/ppblast`

Posters

P1. Characterization of the Bcl-2 family using structure-aided HMM framework
A. Aouacheria, V. Rech de Laval, G. Deléage et C. Combet

P2. Fast and accurate multiple sequence alignment of large and diversified sets of distant homologues
M. Barba, O. Lespinet et B. Labedan

P3. Functional and structural disorder: comparative genomics and genetic interactions distinguish functional roles of disorder
J. Bellay, S. Han, M. Michaut, G. Bader, C. Myers et P. Kim

P4. Computational analysis of the dynamics of logical regulatory graphs
D. Bérenguier, C. Chaouiya, É. Rémy et D. Thieffry

P5. A Rendering Method for Small Molecules up to Macromolecular Systems: HyperBalls Accelerated by Graphics Processors
M. Chavent, A. Vanel, B. Levy, B. Raffin, A. Tek et M. Baaden

P6. Lineage-specific orthologous gene loss and pseudogenisation, automated analysis in Metazoans
J. Dainat, J. Thompson, O. Poch, P. Pontarotti et P. Gouret

P7. Plume: Promoting Economical, Useful and Maintained Software for the Higher Education and the Research Community
C. Dantec et E. Courcelle

P8. The IntAct molecular interaction database and data distribution with PSICQUIC
M. Dumousseau, S. Orchard, B. Aranda, S. Kerrien, J. Khadake, M. Duesbury et H. Hermjakob

P9. IMGT/3Dstructure-DB and tools for immunoglobulins (IG) or antibodies, T cell receptors (TR), MHC, IgSF and MhcSF structural data
F. Ehrenmann et M.-P. Lefranc

P10. MetaBoFlux: a method to analyse flux distribution in metabolic networks
A. Ghozlane, F. Bringaud, F. Jourdan et P. Thébault

P11. IMGT-ONTOLOGY for immunogenetics and immunoinformatics information systems
V. Giudicelli et M.-P. Lefranc

P12. CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources
D. Goudenège, S. Avner, C. Lucchetti-Miganeh et F. Barloy-Hubler

P13. Origin of Phenotypic Specificities in Wine Yeast through a Genomic Approach
C. Guérin, H. Chiapello et P. Nicolas

P14. Structural Analysis of Proteins with Tandem Repeats by Hybrid Approaches
A. Kajava

P15. Evaluating Genome Browsers Using a Software Qualification Method
T. Lacroix, V. Loux, A. Gendrault, J.-F. Gibrat et H. Chiapello

P16. Web Services for MIcrobial Genome Annotation using Data Integration
C. Michotey, L. Legrand, H. Chiapello, V. Loux, A. Gendrault, J.-F. Gibrat et C. Caron

P17. Exact distribution of a pattern in a set of random sequences
G. Nuel, L. Regad, J. Martin et A.-C. Camproux

P18. Influence of the rearrangement rates on the organization of genome transcription
D. Parsons, C. Knibbe et G. Beslon

P19. Meteor – a plateform for quantitative metagenomic profiling of complex ecosystems
N. Pons, J.-M. Batto, S. Kennedy, M. Almeida, F. Boumezbeur, B. Moumen, P. Léonard, E. Le Châtelier, S. Ehrlich et P. Renault

P20. Prediction of patterns of interest from protein primary sequence through structural alphabet
C. Reynès, L. Regad, R. Sabatier et A.-C. Camproux

P21. Scalability of large-scale protein domain inference
C. Rezvoy, D. Kahn et F. Vivien

P22. Counting rna pseudoknotted structures
C. Saule, M. Régnier, J.-M. Steyaert et A. Denise

P23. Computational biology exploration of the enzymatic diversity of an uncharacterised prokaryotic protein family
A. Smith, M. Salanoubat, J. Weissenbach, C. Médigue et D. Vallenet

P24. The Small, Slow and Specialized crispr and Anti-crispr of *Escherichia* and *Salmonella*
M. Touchon et E. Rocha

P25. Differentiation of allelic frequencies analysis identifies short genomic regions with signatures of artificial selection between canine breeds
A. Vaysse, A. Ratnakumar, T. Derrien, K. Lindblad-toh, C. André, M. Webster et C. Hitte

P26. ppalign: posterior probabilities for score-based alignments
S. Wolfsheimer, A. Hartmann et G. Nuel

P27. Imgt/HighV-quest: A High-Throughput System and Web Portal for the Analysis of Rearranged Nucleotide Sequences of Antigen Receptors - High-Throughput Version of imgt/v-quest
E. Alamyar, V. Giudicelli, P. Duroux et M.-P. Lefranc

P28. "GenoVA", a new approach to assess intraspecies genetic variability in complex genomic mixes
M. Almeida, N. Pons, J.-M. Batto, E. Le Châtelier, F. Boumezbeur, N. Sanchez, N. Legravet, C. Delorme, S. Kennedy, S. Ehrlich et P. Renault

P29. DroPNet: Bioinformatics web platform for functional and proteomics data analysis
A. Baillif, M. Agier, E. Mephu-Nguifo et V. Mirouse

P30. Conformational Rearrangements of Lipases Investigated by Molecular Dynamics Simulations
S. Barbe, F. Bordes, A. Marty, L. Mourey, S. Tranier, P. Monsan, M. Remaud-Siméon et I. André

P31. MolliGen 3.0, evolution of a database dedicated to the comparative genomics of mollicutes
A. Barré, C. Lemaître, P. Thébault, A. de Daruvar, A. Blanchard et P. Sirand-Pugnet

P32. UniProt knowledge database and cross references
B. Bely et M. Jesus-Martin

P33. Extension of segm web server to stochastic evolution of tetranucleotides and pentanucleotides
E. Bénard et C. Michel

P34. Renabi grisbi - Grande Infrastructure pour la Bioinformatique
C. Blanchet, C. Gauthey, O. Collin, T. Martin, N. Melab, F. Plewniak, F. Samson, B. Spataro et C. Caron

P35. Reliable identification of hundreds of proteins without peptide fragmentation
P. Bochet, F. Rügheimer, T. Guina, P. Brooks, D. Goodlett, P. Clote et B. Schwikowski

P36. Protic workshop: a bioinformatics environment for proteomics data analysis, validation and integration
J.-P. Bouchet, D. Jeannin, K. Leyre, M. Faurobert, L. Gil, D. Jacob, A. de Daruvar, C. Lalanne, C. Plomion, R. Florès, B. Valot, M. Zivy, O. Langella et J. Joets

P37. An exhaustive mapping method for ngs short reads reveals deficiency of heuristic approaches
F. Boumezbeur, N. Pons, J.-M. Batto, M. Giraud, P. Renault et S. Ehrlich

P38. Exploring the transcriptional response of *Arabidopsis* under stress conditions by a graph-mining approach highlights new insights into key metabolic pathways
F. Boyer, F. Combes, A. Lindlof, J. Bourguignon et Y. Vandenbrouck

P39. Development of a workflow for snp detection with Galaxy
M. Bras et S. Arnoux

P40. Définition de patchs 3d et fouille relationnelle pour la caractérisation et la prédiction de sites d'interactions protéine-protéine
E. Bresso, M. Smaïl-Tabbone et M.-D. Devignes

P41. Mining sequence similarity and microsynteny for functional inference
V. Calderon, R. Barriot, S. de Bentzmann, Y. Quentin et G. Fichant

P42. The EvolScope project : an extension of the MicroScope platform to study the evolution of bacterial polymorphism from high-throughput sequencing data
B. Chane-Woon-Ming, G. Salvignol, D. Vallenet, S. Cruveiller et C. Médigue

P43. Transposition detection using ngs approaches in Asian Rice
C. Chapparo, M. Elbaidouri, N. Picault, O. Panaud et F. Sabot

P44. Mgca: a flexible tool for phylogenomic analysis of prokaryotic genomes
K. Chennen, P. Lechat, É. Hirchaud, R. Cahuzac, P. Dehoux et C. Dauga

P45. A joint experimental and simulation study of aging and protein aggregation in *E. coli*
A.-S. Coquel, A. Demarez, H. Berry et A. Lindner

P46. MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks
L. Cottret, D. Wildridge, F. Vinson, M. Barrett, M.-F. Sagot, H. Charles et F. Jourdan

P47. Gene-rich Domains of the Mammalian Chromatin Display Oscillatory Contact Frequencies
F. Court, J. Miro, C. Braem, M.-N. Le Lay-Taha, A. Brisebarre, F. Atger, T. Gostan, M. Weber, G. Cathala et T. Forné

P48. RNAspace: a web application for ncRNA identification
M.-J. Cros, A. de Monte, J. Mariette, P. Bardou, D. Gautherert, H. Touzet et C. Gaspin

P49. Mgx – Montpellier GenomiX: Plateforme de services en génomique
C. Dantec, J.-P. Desvignes, É. Dubois, A. Brisebarre, G. Baronian, H. Parrinello, D. Séverac, S. Libat et L. Journot

P50. Identification of cis-regulatory elements and functional associations from clusters of genes co-expressed during *Drosophila* zygotic activation
É. Darbo, T. Lecuit, D. Thieffry et J. van Helden

P51. ARPAS: logiciel de gestion de collection de ressources biologiques
S. DEMEY, F. BIMET, E. BÉGAUD, D. CLERMONT, B. CAUDRON, B. PAPIEROK et C. BIZET

P52. SNiPlay, a web application for SNP analysis
A. DEREEPER, S. NICOLAS, L. LECUNFF, R. BACILIERI, A. DOLIGEZ, J.-P. PEROS, P. THIS et M. RUIZ

P53. Discriminating between spurious and significant matches
H. DEVILLERS, M. EL KAROUI et S. SCHBATH

P54. Using R for data management in plant ecophysiology information systems
C. DOMERG, J. FABRE, V. NÈGRE, P. NAUDIN, V. GEORGESCU et A. TIREAU

P55. MeRy-B: a web knowledgebase for the storage, visualization, analysis and annotation of plant metabolomics profiles obtained from NMR
H. DUMAZET, L. GIL, C. DEBORDE, A. MOING, S. BERNILLON, D. ROLIN, A. DE DARUVAR et D. JACOB

P56. The PSICQUIC Interface – a portal into the world of the Interactome
M. DUMOUSSEAU, S. ORCHARD, B. ARANDA, S. KERRIEN et H. HERMJAKOB

P57. Assessing bioinformatics tools for metalloproteins identification: the iron-sulphur proteins case study
J. ESTELLON, S. OLLAGNIER DE CHOUDENS, S. ALVES-CARVALHO, M. FONTECAVE et Y. VANDENBROUCK

P58. Using ontologies for R functions management
J. FABRE, C. DOMERG, É. GENNARI, A. GRANIER, V. NÈGRE, P. NEVEU et A. TIREAU

P59. Logical modelling of the regulatory network controlling the formation of egg appendages in *Drosophila*
A. FAURÉ, B. VREEDE, É. SUCENA et C. CHAOUIYA

P60. Comparison of mapping softwares for next generation sequencing data
J. FAYOLLE, J.-F. GIBRAT, V. LOUX et S. SCHBATH

P61. Towards a multi-scale and formalized representation of protein sequence-structure-function relationships – the nsLTP family as a case of study
C. FLEURY, M.-F. GAUTIER, P. LARMANDE, S. PÉRÈS, F. DE LAMOTTE, F. MOLINA et M. RUIZ

P62. CSPD: a database and search engine for carbonylated proteins
S. GAGNOT, S. DUKAN, C. BROCHIER-ARMANET et E. TALLA

P63. Polymorfind: an automatic pipeline for detecting SNP and indel in sequences of PCR products from heterozygous species
S. GAILLARD, F. FOUCHER et A. PERNET

P64. Mixed-formalism hierarchical modeling and simulation with BioRica
A. GARCIA, D. SHERMAN et R. ASSAR

P65. Logical modelling of MAPK pathways
L. GRIECO, L. CALZONE, A. ZINOVYEV, B. KAHN-PERLES et D. THIEFFRY

P66. Bio++: Object-oriented libraries for sequence analysis, population genetics, molecular evolution and phylogenetics
L. GUÉGUEN, J. DUTHEIL, S. GAILLARD, B. BOUSSAU, G. DUGAS et K. BELKHIR

P67. Prédiction de la structure secondaire des protéines : mise en œuvre optimisée de l'architecture en cascade
Y. Guermeur et F. Thomarat

P68. Chado Controller: un superviseur pour la confidentialité, la qualité et le suivi des annotations
V. Guignon, G. Droc, C. Poiron, J. Lengelle, O. Garsmeur, F.-C. Baurens et S. Sidibé-Bocs

P69. HBVdb: A knowledge database for the *Hepatitis* B Virus
J. Hayer, F. Jadeau, G. Deléage et C. Combet

P70. Validation automatique des sites de phosphorylations par comparaison de scores et intégration d'annotations protéiques sur des données LC/MS-MS
V. Hourdel, O. Jardin-Mathe et M. Vandenbogaert

P71. BYKdb: A database of bacterial tyrosine kinases
F. Jadeau, C. Grangeasse, G. Deléage et C. Combet

P72. Role of geography and languages in shaping population genetic structure
F. Jay, O. François et M. Blum

P73. T-REKS and PRDB: new tools for large scale analysis of protein tandem repeats
J. Jorda et A. Kajava

P74. RNA-seq data analysis provides evidence for a new molecular mechanism generating antisense transcripts in human cells
P. Kapranov, F. Ozsolak, S. Kim, S. Foissac, D. Lipson, C. Hart, S. Roels, C. Borel, S. Antonarakis, A. Monaghan, B. John et P. Milos

P75. Using HOARE logic for constraining parameters of discrete models of gene networks
Z. Khalis, G. Bernot et J.-P. Comet

P76. Control of lipase enantioselectivity by engineering the substrate binding site: An investigation using mixed molecular modelling and robotic-based path planning approaches
V. Lafaquière, S. Barbe, D. Guieysse, J. Cortes, P. Monsan, T. Siméon, M. Remaud-Siméon et I. André

P77. IMGT/LIGMotif: a tool for immunoglobulin and T cell receptor gene identification and description in large genomic sequences
J. Lane, P. Duroux et M.-P. Lefranc

P78. Orylink: a personalized integrated system for plant functional genomic analysis
P. Larmande et G. Droc

P79. MobyleNet: user-centered large spectrum service integration over distributed sites
S. Larroudé, J. Maupetit, H. Ménager, B. Néron, A. Saladin, B. Caudron et P. Tufféry

P80. Un visualisateur dynamique de synténie pour génomes microbiens
P. Lechat, J. Torrent et I. Moszer

P81. TriAnnot V2.0 a friendly web interface for monocot genomic sequences automatic annotation
P. Leroy, H. Sakai, F. Choulet, N. Guilhot, M. Seidel, H. Ohyanagi, A. Bernard, C. Pelegrin, T. Flutre, S. Reboux, M. Alaux, H. Numa, T. Tanaka, N. Amano, K. Mayer, T. Itoh, H. Quesneville et C. Feuillet

P82. Bovine promoter annotation platform for the identification of transcription factor binding sites in genes involved in early pregnancy
M. Leveugle, V. Loux et J.-F. Gibrat

P83. OrthoInspector : comprehensive orthology analysis and visual exploration
B. Linard, J. Thompson, O. Poch et O. Lecompte

P84. Reconstruction and validation of the genome-scale metabolic model of Yarrowia lipolytica iNL750
N. Loira et D. Sherman

P85. Development of knowledge-based system for analysing the effects of single nucleotide polymorphisms on the protein function
T. Luu, N. Nguyen, A. Friedrich, J. Muller, L. Moulinier et O. Poch

P86. Exploring the biodiversity of the world largest ecosystem: BioMarKs project first results and bioinformatics challenges
F. Mahé, S. Audic, R. Christen, J.-M. Claverie, H. Ogata, J. Dolan, B. Edvardsen, W. Kooistra, R. Massana, J. Pawlowski, T. Richards, T. Stoeck et C. De Vargas

P87. A Novel Approach for Comparative Genomics & Annotation Transfer
A. Mancheron, R. Uricaru et É. Rivals

P88. Génolevures, bases de connaissance et annotation des génomes des levures hémiascomycètes
T. Martin, D. Sherman et P. Durrens

P89. 3D Printing Service @ rpbs – mTi
J. Maupetit, B. Villoutreix et P. Tufféry

P90. Logical modelling of drosophila signalling pathways
A. Mbodj, D. Bérenguier, G. Junion, E. Furlong et D. Thieffry

P91. VectorBase, a home for invertebrate vectors of human pathogens
K. Mégy, G. Koscielny et D. Lawson

P92. Assemblage de génomes bactériens séquencés par ngs : comparaison d'outils et choix de paramètres
F. Melchiore, C. Guérin, P. Nicolas et V. Loux

P93. Towards the unbiased prioritarization of Huntington Disease targets using network-based analysis of genome-wide datasets
L. Mesrob, F.-X. Lejeune, C. Bicep, J.-P. Vert et C. Neri

P94. Pepop ou le design de peptides ciblant des épitopes discontinus
V. Moreau, V. Demolombe, G. Lavigne, È. Dupas et C. Granier

P95. TuberGAS: annotation and visualization of the Black Truffle genome
E. Morin, B. Hilselberger, É. Tisserant, B. Brault, B. Noël, B. Porcel, J. Amselem, P. Wincker et F. Martin

P96. Gwas-as: assistance for a thorough evaluation of advanced algorithms dedicated to genome-wide association studies
T. Morisseau, R. Mourad, C. Dina, P. Leray et C. Sinoquet

P97. Investigate Genome Structure and Genes Regulation: a Novel Approach to Identify a Co-Expression Among and Between Groups of Nearby Genes
M. Ouedraogo, S. Lê et F. Lecerf

P98. Anatomy of druggable pockets and associated ligands
S. Pérot et A.-C. Camproux

P99. Digital gene expression data, cross-species conservation and noncoding rna
N. Philippe, F. Rufflé, É. Bou-Samra, A. Boureux, É. Rivals et T. Commes

P100. BioDesc : gestion d'un entrepôt multiformat de descriptions de ressources bioinformatiques
P. Picouet, Z. Doughi, L. Brillet, E. Corre, C. Caron, O. Collin, F. Moreews et X. Bailly

P101. Imgt/mAb-db: the imgt® database for therapeutic monoclonal antibodies
C. Poiron, Y. Wu, C. Ginestoux, F. Ehrenmann, P. Duroux et M.-P. Lefranc

P102. Food-Microbiome, une étude d'éco-génomique appliquée à des écosystèmes fromagers
N. Pons, S. Kennedy, A. Hermet, E. Le Châtelier, M. Almeida, J.-M. Batto, F. Boumezbeur, B. Quinquis, N. Galleron, J.-L. Jany, G. Barbier, Y. Brygoo, C. Delorme, E. Guédon et P. Renault

P103. Abgd, Automatic Barcode Gap Discovery
N. Puillandre, A. Lambert, S. Brouillet et G. Achaz

P104. PALmapper: Fast and Accurate Spliced Alignments of rna-seq Reads
G. Raetsch, G. Jean, A. Kahles, S. Sonnenburg, F. De Bona, K. Schneeberger, J. Hagmann et D. Weigel

P105. Sequence analysis of the proteins involved in CaCO3 biomineralization
P. Ramos-Silva, F. Marin, G. Deléage et C. Combet

P106. Hierarchical classification of helical distortions related to proline
J. Rey, J. Devillé et M. Chabbert

P107. Stratégie de recherche et d'annotations de nouvelles synthétases non-ribosomiales à partir de génomes bactériens
Z. Saci, M. Pupin, J. Deravel, F. Krier, S. Caboche, P. Jacques et V. Leclère

P108. Rna locally optimal secondary structures
A. Saffarian, M. Giraud et H. Touzet

P109. Functional prediction in the scope of large-scale multi-class learning
R. Saidi, S. Aridhi, M. Agier, G. Bronner, D. Debroas, L. D'Orazio, F. Enault, S. Guillaume et E. Mephu-Nguifo

P110. Comparing graph-based representations of protein for mining purposes
R. Saidi, M. Maddouri et E. Mephu-Nguifo

P111. PotChips: a new method for chip-chip data analysis
F. Salipante, C. Reynès, L. Journot, C. Dantec et R. Sabatier

P112. Processus d'analyse statistique pour la découverte de biomarqueurs en diagnostic médical
N. Salvetat, È. Dupas, K. Kaminski et F. Molina

P113. Estimating the size of the *S. cerevisiae* interactome
L. Sambourg et N. Thierry-Mieg

P114. Development and optimization of metagenomic analyses
L. Siegwald, F. Texier et C. Hubans-Pierlot

P115. BioInformatic analyses of sex-determination in Tilapia (*Oreochromis spp*)
L. Soler, M. Conte, T. Kocher, J.-F. Baroiller, H. D'Cotta et I. Mougenot

P116. Pipeline for the pre-processing of Illumina reads
É. Souche, S. Brisse et I. Moszer

P117. A high throughput multi-technological Research Information Management System for the Joomla cms: djeen
O. Stahl, A. Guille, P. Finetti, P. Grenot et G. Bidaut

P118. Proteoscan-db: an open-source pipeline for automatic validation of phosphopeptides from cid ms spectra
M. Tauzin, A. Lermine, M. Rossignol et C. Nespoulous

P119. EuPathDomains: The Divergent Domain Database for Eukaryotic Pathogens
N. Terrapon, A. Ghouila, O. Gascuel, F. Guerfali, D. Laouini, É. Maréchal et L. Bréhélin

P120. Tfm-Explorer: mining cis-regulatory regions in genomes
L. Tonon, H. Touzet et J.-S. Varré

P121. PhyloWeb : A dynamic web viewer for microbial population genetics
J. Torrent, G. Guigon et I. Moszer

P122. Can we link metagenome gene content and iron supply in the ocean ?
È. Toulza, A. Tagliabue, L. Bopp, S. Blain et G. Piganeau

P123. A platform for real-time control of gene expression
J. Uhlendorf, S. Bottani, F. Fages, P. Hersen et G. Batt

P124. Oasys : Un outil dédié à la visualisation et à l'exploration des données de biopuces
A.-S. Valin, L. Marisa, L. Vescovo, È. Thomas, M. Guedj, R. Schiappa, J. Laffaire, F. Petel et A. de Reynies

P125. ISsaga, a platform for identification and semi-automatic annotation of prokaryotic insertion sequences
A. Varani, P. Siguier, É. Gourbeyre, V. Charneau et M. Chandler

P126. Wheat and barley data in Gnpis, The urgi information system
D. Verdelet, M. Alaux, N. Mohellibi, S. Durand, E. Kimmel, I. Lyuten, S. Reboux, D. Steinbach et H. Quesneville

P127. Topological characteristics of the functionalization process for duplicated genes in ppi networks of *Arabidopsis thaliana*
R. Zaag, É. Birmelé et C. Rizzon

P128. S-mart: how to handle your rna-Seq data?
M. Zytnicki et H. Quesneville

Liste des conférences invitées

Liste des présentations longues

Liste des présentations courtes

Liste des contributeurs

- W —

- Z —

Journées Ouvertes de Biologie, Informatique et Mathématiques

Montpellier, 7 - 9 septembre 2010

La conférence JOBIM est née il y a 10 ans à Montpellier, où elle revient cette année. C'est un lieu de rencontre ouvert à toutes les personnes travaillant aux frontières de la biologie, de l'informatique, des mathématiques et de la physique, afin de favoriser les échanges scientifiques et d'encourager l'expression des jeunes chercheurs. Les grands thèmes sont liés à la génomique, la bioinformatique structurale, la biologie des systèmes et l'analyse des données d'expression, l'évolution et la phylogénie, les bases de données et de connaissances, l'algorithmique et la modélisation, en particulier issue des probabilités et des statistiques. Mais la discipline se renouvelle et voit de nouveaux champs s'ouvrir, par exemple en analyse d'images, en génétique des populations ou du côté de l'écoinformatique. Elle bénéficie de données toujours plus abondantes et diverses, notamment de séquences grâce à l'amélioration spectaculaire des techniques de séquençage. Ces données à grande échelle permettent de répondre à de nouvelles questions, liées à l'épigénétique par exemple, mais elles imposent aussi de revoir les méthodes et les techniques.

Nous avons reçu cette année 66 soumissions, 16 ont été retenues pour des présentations longues et 26 pour des présentations courtes, auxquelles s'ajoutent les conférences invitées de Raphaël Guérois, Jean-Christophe Olivo-Marin, Luis Quintana-Murci, Sven Rahmann, Jörg Stelling et Pierre Taberlet. Ces actes contiennent les articles associés à l'ensemble de ces présentations, ainsi que la liste des quelques 130 posters qui seront affichés et discutés lors de la conférence.