JOBIM

JOURNÉES
OUVERTES
BIOLOGIE
INFORMATIQUE
MATHÉMATIQUES

TOULOUSE 2013

Volume 2/2 **RÉSUMÉS COURTS [affiches]**

[1-4 JUILLET 2013]

JOBIM TOULOUSE 2013

JOURNÉES OUVERTES BIOLOGIE INFORMATIQUE MATHÉMATIQUES

Journées
Ouvertes
Biologie
Informatique
Mathématiques

Toulouse, 1- 4 Juillet 2013

Editeurs
Christine Gaspin
Nic Lindley
Cédric Notredame

Préface

Toulouse accueille la quatorzième édition des Journées Ouvertes en Biologie, Informatique et Mathématiques. Depuis 13 ans maintenant, cette conférence rassemble, dans le cadre d'un rendez-vous annuel de partages et d'échanges, la communauté francophone bio(informatique/statistique/mathématique). Comme chaque année, cette conférence est placée sous l'égide de la Société Française de BioInformatique.

Les progrès fulgurants réalisés ces dernières années dans les technologies d'acquisition de données sur le Vivant renouvellent nos questions de recherche en Biologie, Informatique et Mathématiques, générant de nouveaux défis à la fois pour faire face à la masse de données produites, mais aussi pour répondre à la complexité et à la diversité des questions posées. Cette année, nous accueillons huit conférenciers invités de renommée internationale dans notre communauté : SIMON ANDERS, CHRISTINE BRUN, LAURENT DURET, JEAN-LOUP FAULON, RODERIC GUIGO et PEDRO MENDES. Nous les remercions chaleureusement d'avoir accepté de participer à la réussite de ces journées en nous exposant quelques unes de leurs réussites scientifiques récentes face à ces défis.

Nous remercions aussi l'ensemble des relecteurs sollicités dans le comité de programme (et au-delà) pour leur travail important sur les 144 soumissions reçues. Nous espérons que leurs commentaires auront aidé le plus grand nombre à améliorer la qualité des contributions : 37 ont été sélectionnées pour une présentation orale et 97 seront présentées sous forme d'affiches pour être discutées tout au long de ces journées.

Cette année, nous avons pu financer la participation de 4 jeunes chercheurs français réalisant un séjour postdoctoral à l'étranger. Nous espérons que ces journées contribueront à favoriser leur intégration future dans notre communauté. Deux jeunes scientifiques se verront récompensés par l'attribution des prix de la meilleure présentation orale et du meilleur poster. Par ces deux prix nous souhaitons reconnaître des travaux prometteurs et encourager les jeunes chercheurs à poursuivre leurs recherches dans le domaine de la bioinformatique.

Un grand merci à nos partenaires industriels et institutionnels, aux collectivités territoriales locales, à tous les membres des comités de programme et d'organisation ainsi qu'à l'ensemble des bénévoles qui ont largement œuvré à la réussite de JOBIM 2013.

Benvenguts à totas e totes a Tolosa ...et très bonne conférence à tous !!!

Pour le comité de programme
Christine Gaspin, Inra Toulouse
Nic Lindley, Cnrs/Insa GenoToul Toulouse
Cédric Notredame, CRG Barcelone

Pour le comité d'organisation
Christine Cierco-Ayrolles, Inra Toulouse
Monique Falières, Inra GenoToul Toulouse
Maria Martinez, Inserm Toulouse

Comité d'organisation

Fabienne Ayrignac
Hélène Chiapello
Christine Cierco-Ayrolles
Clotilde Claudel
Monique Falières
Thomas Faraut
Sylvain Foissac
Sandra Fuentes
Claire Hoede
Nathalie Julliand
Christophe Klopp
Didier Laborie
Christelle Labruyère
Nic Lindley
Jérôme Mariette
Maria Martinez
Annick Moisan
Céline Noirot
Thomas Schiex
Matthieu Vignes

Avec l'aide des bénévoles de l'unité Mathématiques et Informatique Appliquées de Toulouse

Comité de programme

Gilles Bernot
Vincent Berry
Laurent Bréhélin
Céline Brochier-Armanet
Anne-Claude Camproux
Hélène Chiapello
Eric Coissac
François Coste
Ludovic Cottret
Olivier Cuvier
Hidde De Jong
Sébastien Duplessis
Guillaume Filion
Christine Froidevaux
Olivier Gascuel
Christine Gaspin
Mathieu Giraud
Fabrice Jossinet
Fabien Jourdan
Béatrice Laurent
Dominique Lavenier
Stéphane Le Crom
Claire Lemaitre
Nic Lindley
Juliette Martin
Claudine Médigue
Yves Moreau
Macha Nikolski
Céline Noirot
Cédric Notredame
Guy Perrière
Pierre Peterlongo
Olivier Poch
Yann Ponty
Yves Quentin
Eric Rivals
Hugues Roest Crollius
Irena Rusu
Magali San Cristobal
Sophie Schbath
Hervé Seitz
Thomas Simonson
Dominique Tessier
Denis Thieffry
Patricia Thébault
Hélène Touzet
Pierre Tuffery
Jacques van Helden
Alain Viari

Relecteurs additionnels

Wassim Abou-Jaoudé
Samuel Blanquart
Sarah Cohen-Boulakia
Maude Guillier
Bernard Labedan
François Le Fevre
Alban Mancheron
Jacques Nicolas
Michel Petitjean
Mikaël Salson
Hayssam Soueidan
Morgane Thomas-Chollier

Types de contributions

En sus des résumés de nos 8 conférenciers invités, les contributions présentées dans ce recueil sont de trois types :

- **Articles originaux (8-10 pages)** : réservés aux résultats originaux non publiés par ailleurs. Les contributions sélectionnées dans cette catégorie donnent lieu à une présentation orale.
- **Résumés étendus (2-8 pages)** : ces contributions proposent des résultats récents, éventuellement déjà soumis ou acceptés ailleurs. Les contributions sélectionnées dans cette catégorie donnent lieu à une présentation orale.
- **Résumés courts (1-2 pages)** : ces contributions concernent des travaux en cours, publiés ou non. Les contributions sélectionnées dans cette catégorie donnent lieu à une présentation sous forme d'affiche.

Table des matières

(par numéro de poster)

Computational phenotype prediction of ionizing-radiation-resistant bacteria with a multiple-instance learning model

Sabeur ARIDHI[124], Haitham SGHAIER[3], Mondher MADDOURI[4] and Engelbert MEPHU NGUIFO[12]

[1] LIMOS - UBP - Clermont University, BP 10125, 63173, Clermont Ferrand, France
[2] CNRS, UMR 6158, LIMOS, F-63173 Aubiere, France
[3] Research Unit UR04CNSTN01 ”Medical and Agricultural Applications of Nuclear Techniques” - National Center for Nuclear Sciences and Technology (CNSTN), Sidi Thabet Technopark, Ariana 2020, Tunisia
[4] LIPAH - Faculty of sciences of Tunis - University of Tunis El Manar, 1060, Tunis, Tunisia

Abstract *Ionizing-radiation-resistant bacteria (IRRB) are important in biotechnology. In silico methods are unavailable for the purpose of phenotypic prediction and genotype-phenotype relationship discovery. We analyzed basal DNA repair proteins of most known proteomes sequences of IRRB and ionizing-radiation-sensitive bacteria (IRSB) in order to learn a classifier that correctly predicts unseen bacteria. In this paper, we formulated the problem of predicting IRRB as a multiple-instance learning (MIL) problem and we proposed a novel approach for predicting IRRB. We used a local alignment technique to measure the similarity between protein sequences to predict ionizing-radiation-resistant bacteria. First experimental results are satisfactory.*

Keywords Phenotypic prediction, ionizing-radiation-resistant bacteria, ionizing-radiation-sensitive bacteria, protein sequences, multiple-instance learning

1 Background

Nuclear waste contains a variety of toxic and radioactive substances. The bioremediation of these wastes with pertinent bacteria and low cost is a challenging problem [1,2]. The use of ionizing-radiation-resistant bacteria (IRRB) for the treatment of these radioactive wastes is determined by their surprising capacity of adaptation to radionuclides and to a variety of toxic molecules. To date, genomic databases indicate the presence of thousands of genome projects. However, only a few computational works are available for the purpose of phenotypic prediction discovery that rapidly determines useful genomes for the bioremediation of radioactive wastes [2].

A main idea in this context is that resistance to ionizing radiation and tolerance of desiccation are two complex phenotypes, and suggest that protection and repair mechanisms are complementary in IRRB. In addition, it seems that the shared ability of IRRB to survive the damaging effects of ionizing radiation and desiccation is the result of basal DNA repair pathways and that basal DNA repair proteins in IRRB, unlike many of their orthologs in IRSB, present a strong ability to effectively repairs damage incurred to DNA.

In this work, we study the basal DNA repair protein of IRRB and IRSB to solve the problem of phenotypic prediction in IRRB. Thus, we consider that each studied bacterium is represented by a set of DNA repair proteins. Due to this fact, we formalize the problem of phenotypic prediction in IRRB as a multiple instance learning problem (MIL) in which bacteria represent bags and repair proteins of each bacterium represent instances.

Many multiple instance learning algorithms have been developed to solve several problems such as predicting types of Protein-Protein Interactions (PPI) [3] and drug activity prediction [4], mainly including Diverse Density [5], Citation-kNN and Bayesian-kNN [6], MI SVMs [7] and HyDR-MI [8]. Diverse Density (DD) was proposed in [5] as a general framework for solving multi-instance learning problems. The main idea of DD approach is to find a concept point in the feature space that are close to at least one instance from every positive bag and meanwhile far away from instances in negative bags. The optimal concept point is defined as the one with the maximum diversity density, which is a measure of how many different positive bags have instances near the point, and how far the negative instances are away from that point. In [6], the minimum Hausdorff distance was used as the bag-level distance metric, defined as the shortest distance between any two instances from each bag. Using this bag-level distance, we can predict the label of an unseen bag using the *k-NN* algorithm. In [7], the authors proposed the algorithm *mi-SVM* to modify Support Vector Machines. The algorithm *mi-SVM* explicitly treats the label instance labels as unobserved hidden variables subject to constraints defined by their bag labels. The goal is to maximize the usual instance margin jointly over the unknown instance labels and a linear or kernelized discriminant function. In [8], the authors proposed a feature subset selection method for MIL algorithms called HyDR-MI (hybrid dimensionality reduction method for multiple instance learning). The hybrid consists of the filter component based on an extension of the ReliefF algorithm [9] developed for working with MIL and the wrapper component based on a genetic algorithm that optimizes the search for the best feature subset from a reduced set of features, output by the filter component.

The above cited algorithms use an attribute-value format to represent their data. A most used approach to represent protein sequences in an attribute-value format is to extract motifs that can serve as attributes. Appropriately chosen sequence motifs may reduce noise in the data and indicate active regions of the protein. A protein can then be represented as a set of motifs [10,11] or as a vector in a vector space spanned by these motifs [12]. However, the use of this technique is not suitable in the context of phenotypic prediction of IRRB. This is due to the fact that the set of proteins of each bag must be represented (in the attribute-value format) with the same set of attributes which is possible only if all extracted motifs from the different bag of proteins are putting together as a unique set of motifs. As the different bags of proteins are processed disjointly, it is necessary to design a novel approach for such case.

In this paper, we propose a MIL approach for predicting IRRB using proteins implicated in basal DNA repair in IRRB. We used a local alignment technique to measure the similarity between protein sequences of the studied bacteria to predict ionizing-radiation-resistant bacteria. To the best of our knowledge, this is the first work which proposes an *in silico* approach for phenotypic prediction in IRRB.

The remainder of this paper is organized as follows. Section 2 presents the materials and methods used in the our study. In Section 3, we describe our experimental techniques and we discuss the obtained results. Concluding points make the body of Section 4.

2 Materials and Methods

2.1 Terminology and problem formulation

The task of multiple instance learning (MIL) was coined by Dietterich et al. [13] when they were investigating the problem of drug activity prediction. In multiple-instance learning, the training set is composed of n labeled bags. Each bag in the training set contains k instances and have a bag label $y_i \in \{-1, +1\}$. We notice that instances of each bag have labels $y_{ij} \in \{-1, +1\}$, but these values are not known during training. The most common assumption in this field is that a bag is labeled positive if at least one of its instances is positive, which can be expressed as follows:

$$y_i = \max_j(y_{ij}). \tag{1}$$

The task of MIL is to learn a classifier from the training set that correctly predicts unseen bags. Although MIL is quite similar to traditional supervised learning, the main difference between the two approaches can be found in the class labels provided by the data. According to the specification given by Dietterich et al. [13], in a traditional setting of machine learning, an object m is represented by a feature vector (an instance) which is associated to a label. However, in a multiple instance setting, each object m may have k various instances denoted $m_1, m_2, \cdots, m_k$. The difference between the traditional setting of machine learning and the multiple instance learning setting can be represented clearly in Fig. 1 where the difference between the input objects is shown.

In our work, we are interested to a specific bacteria family with high radioresistance to ionizing radiation and tolerance of desiccation. This family contains a set of bacteria. Let $DB = \{X_1, \ldots, X_n\}$ be a bacteria

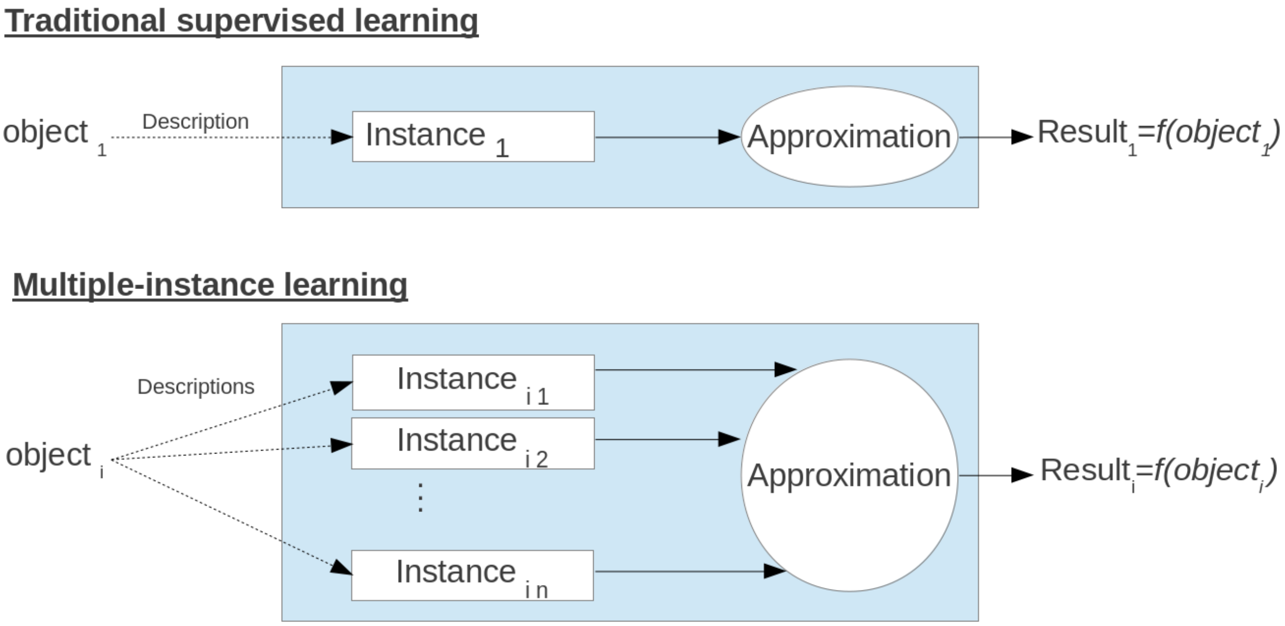


Figure 1. Differences between traditional supervised learning and multiple instance learning.

database. Each bacterium in the database is represented by a set of proteins $X_i = \{p_{i1}, \cdots, p_{ik}\}$ and belongs to a class label y_i with $y_i = \{IRRB, IRSB\}$. The problem of phenotypic prediction of IRRB can be seen as a MIL problem in which bacteria represent bags, and basal DNA repair proteins of each bacterium represent instances.

The problem investigated in this work is to learn a multiple-instance classifier in this setting. Given a query bacterium $Q = \{p_1, \cdots, p_k\}$, the classifier must use primary structures of basal DNA repair proteins in Q and in each bag of DB to predict the label of Q.

2.2 MIL-ALIGN algorithm

Based on the formalization, we propose the MIL-ALIGN algorithm allowing to predict ionizing-radiation-resistant bacteria. The proposed algorithm focuses on discriminating bags by the use of local alignment technique to measure the similarity between each protein sequence in the query bag and corresponding protein sequence in the different bags of the learning database.

In MIL-ALIGN algorithm we use the following variables for input data and for accumulating data during the execution of the algorithm:
- the variable Q: corresponds to the query bag (the query bacterium) which is a vector of protein sequences.
- the variable DB: corresponds to the bacteria database.
- the variable S: corresponds to a matrix used to store alignment score vectors.

Algorithm 1 MIL-ALIGN

```
Require: Learning database $DB = \{(X_1, y_1), \cdots, (X_n, y_n)\}$, Query $Q = \{p_1, \cdots, p_k\}$
Ensure: Prediction result $R$
for all $p_i \in Q$ do
    for all $X_j$ do
        $s_{ij} \leftarrow LocalAlignment(p_i, p_{ji})$    //$X_j = \{p_{j1}, \cdots, p_{jk}\}$ and $p_{ji}$ is the protein number $i$ of bacterium $X_j$
        $S_{ij} \leftarrow s_{ij}$
    end for
end for
$R \leftarrow Aggregate(S)$
return $R$
```

Informally, the algorithm works as follows (see Algorithm 1):

1. For each protein sequence p_i in the query bag Q, MIL-ALIGN computes the corresponding alignment scores (line 1 to 6).
2. Group alignment scores of all protein sequences of query bacterium into a matrix S (line 4). Line i of S corresponds to a score vector of protein p_i against all proteins p_{ji} of X_j with $1 \leq j \leq n$. Element S_{ij} corresponds to the alignment score of protein p_i with protein p_{ji} of bacterium X_j.
3. Apply an aggregation method to S in order to compute the final prediction result R (line 8). A query bacterium is predicted as IRRB (respectively IRSB) if the aggregation result of similarity scores of its proteins against associated proteins in the learning database is IRRB (respectively IRSB).

2.3 Experimental environment

Information on complete and ongoing IRRB genome sequencing projects was obtained from the GOLD database [14]. We initiated our analyses by retrieving orthologous proteins implicated in basal DNA repair in IRRB with fully sequenced genomes. Table 1 presents the used IRRB and IRSB.

Table 1. IRRB and IRSB

Bacterium	**Phenotype**
Acinetobacter radioresistens SH164 *Kineococcus radiotolerans* SRS30216 *Methylobacterium radiotolerans* JCM 2831 *Deinococcus maricopensis* DSM 21211 *Gemmata obscuriglobus* UQM 2246 *Deinococcus proteolyticus* MRP *Truepera radiovictrix* DSM 17093 *Acinetobacter radioresistens* SK82	IRRB
Escherichia coli OP50 *Neisseria gonorrhoeae* MS11 *Neisseria gonorrhoeae* PID1 *Neisseria gonorrhoeae* DGI18 *Pseudomonas putida* S16 *Thermus thermophilus* SG0.5JP17-16	IRSB

For our experiments, we constructed a training set containing 14 bags (8 IRRB and 6 IRSB). Each bag contains at most 30 instances which correspond to proteins implicated in basal DNA repair in IRRB (see Table 2). Protein sequences were downloaded from the FTP site of the curated database SwissProt [1].

3 Results and Discussion

3.1 Experimental Techniques

The computations were carried out on a duo CPU 2.86 GHz PC with 2 GB memory, operating on Ubuntu Linux. In the classification process, we used the Leave-One-Out (LOO) technique [15] also known as *jackknife test*. For each dataset (comprising n bags), only one bag is kept for the test and the remaining part is used for the training. This action is repeated n times. In our context, the leave-one-out is considered to be the most objective test technique compared to the other ones (i.e., hold-out, n-cross-validation) as our training set contains a small number of bacteria.

For our tests, we used the BLAST tool [16] for computing local alignments. We implemented two aggregation methods to be used with MIL-ALIGN: the *Sum of Maximum Scores* method and the *Weighted Average of Maximum Scores* method.

Sum of Maximum Scores (SMS). For each protein in the query bacterium, we traverse the corresponding line of S which contains the obtained scores against all other bacteria of the training database. The *SMS* method selects the maximum score among the alignments scores against IRRB bacteria (which we call max_R)

[1] http://www.uniprot.org/downloads

Table 2. Replication, repair, and recombination proteins related to ionizing-radiation-resistant bacteria

ID	Protein	Function
1	DNA polymerase III, α subunit	DNA polymerase
2	DNA polymerase III, ϵ subunit	
3	Putative DNA polymerase III, δ subunit	
4	DNA-directed DNA polymerase	
5	DNA polymerase III, τ/γ subunit	
6	Single-stranded DNA-binding protein	Replication complex
7	Replicative DNA helicase	
8	DNA primase	
9	DNA gyrase, subunit B	
10	DNA topoisomerase I	
11	DNA gyrase, subunit A	
12	Smf proteins	Other DNA-associated proteins
13	Endonuclease III	
14	Holliday junction resolvase	
15	Formamidopyrimidine-DNA glycosylase	
16	Holliday junction DNA helicase	
17	RecF protein	
18	DNA repair protein	
19	Holliday junction binding protein	
20	Excinuclease ABC, subunit C	
21	Transcription-repair coupling factor	
22	Excinuclease ABC, subunit A	
23	DNA helicase II	
24	DNA helicase RecG	
25	Exonuclease SbcC	
26	Ribonuclease HII	
27	Excinuclease ABC, subunit B	
28	A/G-specific adenine glycosylase	
29	RecA protein	
30	DNA-3-methyladenine glycosidase II, putative	

and the maximum score among the scores of alignments against IRSB bacteria (which we call max_S). It then compares these scores. If max_R is greater than max_S, it adds max_R to the total score of IRRB (which we call $total_R(S)$). Otherwise, it adds max_S to the total score of IRSB (which we call $total_R(S)$). When all the selected proteins were traversed, the SMS method compares the total scores of IRRB and IRSB. If $total_R(S)$ is greater than $total_S(S)$, prediction refers IRRB. Otherwise, prediction refers IRSB.

Below, we formally define the SMS method:

$$\text{SMS}(S) = \begin{cases} IRRB, & \text{if } total_R(S) \geq total_S(S), \\ IRSB, & \text{otherwise,} \end{cases}$$

where

– $total_R(S) = \sum_{i=1}^{n} \max_{1\leq j\leq k} \; S_{ij}$ such that $y_j = IRRB$, and

– $total_S(S) = \sum_{i=1}^{n} \max_{1\leq j\leq k} \; S_{ij}$ such that $y_j = IRSB$.

Weighted Average of Maximum Scores (WAMS). With the *WAMS* method, each protein p_i has a given weight w_i. For each protein in the query bacterium, we traverse the corresponding line of S which contains the obtained scores against all other bacteria of the training database. The *WAMS* method selects the maximum score among the scores of alignments against IRRB bacteria (which we call $max_R(S)$) and the maximum score among the scores of alignments against IRSB bacteria (which we call $max_S(S)$). It then compares these scores. If the $max_R(S)$ is greater than $max_S(S)$, it adds $max_R(S)$ multiplied by the weight of the protein to the total score of IRRB and it increments the number of IRRB having a max score. Otherwise, it adds $max_S(S)$ multiplied by the weight of the protein to the total score of IRSB and it increments the number of IRSB having a max score. When all the selected proteins were traversed, we compare the average of total scores of IRRB (which we called $avg_R(S)$) and the average of total scores of IRSB (which we called $avg_S(S)$). If $avg_R(S)$ is greater than $avg_S(S)$, prediction refers IRRB. Otherwise, prediction refers IRSB.

Below, we formally define the WAMS method:

$$\text{WAMS}(S) = \begin{cases} IRRB, & \text{if } avg_R(S) \geq avg_S(S), \\ IRSB, & \text{otherwise,} \end{cases}$$

where

- $avg_R(S) = total_R(S)/num_R$, and
- $avg_S(S) = total_S(S)/num_S$,

and

- $total_R(S) = \sum_{i=1}^{n} \max_{1\leq j\leq k}\ S_{ij} \cdot w_i$ such that $y_j = IRRB$, and
- $total_S(S) = \sum_{i=1}^{n} \max_{1\leq j\leq k}\ S_{ij} \cdot w_i$ such that $y_j = IRSB$,

where w_i is the weight of the protein p_i.

3.2 Results

In order to study the importance of considering the problem of predicting ionizing-radiation-resistant bacteria as a multiple instance learning problem, we present in Table 3 MIL-ALIGN results with one instance for each bag, i.e., each bacterium is represented by one protein sequence. The LOO-based evaluation technique was used to generate the presented results. As shown in Table 3, we conducted our experiments on only 22 proteins. This is due to the fact that experiments on proteins which are not expressed at least for one IRRB bacterium and for one IRSB bacterium were not conducted.

Table 3. Prediction results with the traditional setting of machine learning

Protein	Learning database		Accuracy	Sensitivity	Specificity
	IRRB	IRSB			
DNA primase	8	6	85.7 %	87.5 %	83.3 %
Replicative DNA helicase	8	6	78.5 %	85.7 %	71.4 %
DNA topoisomerase I	8	6	78.5 %	85.7 %	71.4 %
DNA gyrase, subunit A	8	6	71.4 %	75 %	66.6 %
Endonuclease III	8	6	71.4 %	70 %	75 %
Formamidopyrimidine-DNA glycosylase	8	6	71.4 %	75 %	66.6 %
RecA Protein	8	6	64.2 %	66.6 %	60 %
DNA polymerase III, α subunit	8	6	57 %	66.6 %	55.5 %
Excinuclease ABC, subunit A	8	4	75 %	87.5 %	60 %
DNA helicase RecG	5	6	90.9 %	83.3 %	100 %
Excinuclease ABC, subunit C	6	5	81.8 %	100 %	71.4 %
Transcription-repair coupling factor	6	5	72.7 %	71.4 %	75 %
DNA polymerase III, τ/γ subunit	6	5	72.7 %	80 %	66.6 %
DNA gyrase, subunit B	5	6	63.6 %	60 %	66.6 %
Holliday junction resolvase	4	6	70 %	66.6 %	71.4 %
DNA polymerase III, ϵ subunit	6	3	77.7 %	83.3 %	66.6 %
Excinuclease ABC, subunit B	6	3	44.4 %	66.6 %	33.3 %
RecF protein	5	3	75 %	80 %	66.6 %
A/G-specific adenine glycosylase	7	1	75 %	85.7 %	0 %
Single-stranded DNA-binding protein	6	2	50 %	66.6 %	0 %
Ribonuclease HII	2	5	85.7 %	66.6 %	100 %
DNA-directed DNA polymerase	4	1	60 %	75 %	0 %

In order to study the incorrectly classified bacteria, we computed for each bacterium in the learning database, the percentage of failed predictions (see Table 4).

As shown in Table 4, some bacteria present high rates of failed predictions. This means that MIL-ALIGN fails to correctly predict the phenotype of those bacteria with most proteins. On the other hand, the results illustrated in Table 4 may help to understand some characteristics of the studied bacteria. For example, the *Thermus thermophilus* SG0.5JP17-16 bacterium presents a high rate of failed predictions (68.18 %). It mean

Table 4. Percentage of failed predictions

Phenotype	Bacterium	Rate of failed predictions (%)
IRRB	Acinetobacter radioresistens SH164	15
	Kineococcus radiotolerans SRS30216	33.33
	Methylobacterium radiotolerans JCM 2831	77.77
	Deinococcus maricopensis DSM 21211	0
	Gemmata obscuriglobus UQM 2246	47.05
	Deinococcus proteolyticus MRP	5.88
	Truepera radiovictrix DSM 17093	27.77
	Acinetobacter radioresistens SK82	11.11
IRSB	*Escherichia coli* OP50	20
	Neisseria gonorrhoeae MS11	6.25
	Neisseria gonorrhoeae PID1	0
	Neisseria gonorrhoeae DGI18	0
	Pseudomonas putida S16	47.61
	Thermus thermophilus SG0.5JP17-16	83.33

that in most cases, *Thermus thermophilus* SG0.5JP17-16 is predicted as IRRB. This result shows that *Thermus thermophilus* SG0.5JP17-16 might allow a strong ability for DNA protection and repair mechanisms and confirm the *in vitro* results presented in [17] and [18].

Table 5 presents the experimental results of MIL-ALIGN using the whole set of proteins to represent the studied bacteria. For each aggregation method, we present the accuracy of MIL-ALIGN, the sensitivity and the specificity. We notice that the WAMS aggregation method was used with equally weighted proteins.

Table 5. Experimental results of MIL-ALIGN with LOO-based evaluation technique

Aggregation method	Accuracy	Sensitivity	Specificity
SMS	57.1 %	100 %	50 %
WAMS	85.7 %	80 %	100 %

Results in Table 3 show that the use of our algorithm with just one instance for each bag in the learning database allow good accuracy values especially with some specific proteins. For example, the *DNA helicase RecG* protein allows an accuracy value that exceed the one produced by MIL-ALIGN using the whole set of protein sequences (see Table 5). However, almost all results were generated without the whole set of bacteria. In fact, when a protein is not expressed in a specific bacterium, we do not use the bacterium in the learning database. For example, the protein *DNA helicase RecG* is expressed for only 11 bacteria (5 IRRB and 6 IRSB) from the set of 14 bacteria used to generate results presented in Table 5.

We notice that the use of the whole set of proteins to represent the studied bacteria allows good accuracy accompanied by a high values of sensitivity and specificity especially with the WAMS aggregation method. As mentioned in Table 5, with the WAMS aggregation method, we have 85.7 % of accuracy, 80 % of sensitivity and 100 % of specificity. We do not exceed these values in almost all the cases presented in Table 3. The 100 % of specificity presented by MIL-ALIGN with the whole set of proteins to represent the studied bacteria shows the ability of MIL-ALIGN to identify negative bags (IRSB bacteria).

4 Conclusions

In this paper, we addressed the issue of predicting ionizing-radiation-resistant bacteria (IRRB). We have considered that this problem is a multiple-instance learning problem in which bacteria represent bags and repair proteins of each bacterium represent instances. We have formulated the studied problem and described our proposed algorithm MIL-ALIGN for phenotype prediction in the case of IRRB. By running experiments on a real dataset, we have shown that first results of MIL-ALIGN are satisfactory.

In the future work, we will study the performance of the proposed approach to improve its efficiency. Also, we will study the use of a priori knowledge to improve the efficiency of our algorithm. This a priori knowledge can be used to assign weights to proteins during the learning step of our approach. A notable interest will be dedicated to the study of other proteins that can be involved to the high resistance of IRRB to the ionizing radiations and desiccation. In fact, many antioxidant enzymes may play important roles in scavenging free radicals caused by irradiation [19].

5 Acknowledgements

This work was supported by the French-Tunisian DGRST/CNRS project IRRB 11/R 14-09.

References

[1] M. V. Omelchenko, Y. I. Wolf, E. K. Gaidamakova, V. Y. Matrosova, A. Vasilenko, M. Zhai, M. J. Daly, E. V. Koonin, K. S. Makarova, Comparative genomics of Thermus thermophilus and Deinococcus radiodurans: divergent routes of adaptation to thermophily and radiation resistance., BMC Evol Biol 5 (2005) 57.

[2] H. Sghaier, K. Ghedira, A. Benkahla, I. Barkallah, Basal dna repair machinery is subject to positive selection in ionizing-radiation-resistant bacteria, BMC Genomics 9 (1) (2008) 297.

[3] H. Yamakawa, K. Maruhashi, Y. Nakao, Predicting types of protein-protein interactions using a multiple-instance learning model, in: T. Washio, K. Satoh, H. Takeda, A. Inokuchi (Eds.), New Frontiers in Artificial Intelligence, Vol. 4384 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2007, pp. 42–53.

[4] G. Fu, X. Nan, H. Liu, R. Y. Patel, P. R. Daga, Y. Chen, D. Wilkins, R. J. Doerksen, Implementation of multiple-instance learning in drug activity prediction, BMC Bioinformatics 13 (S-15) (2012) S3.

[5] O. Maron, T. L. Pérez, A Framework for Multiple-Instance Learning, in: M. I. Jordan, M. J. Kearns, S. A. Solla (Eds.), Advances in Neural Information Processing Systems, Vol. 10, The MIT Press, Cambridge, MA, 1998, pp. 570–576.

[6] J. Wang, , J.-D. Zucker, Solving the multiple-instance problem: A lazy learning approach, in: In Proc. 17th International Conf. on Machine Learning, Morgan Kaufmann, 2000, pp. 1119–1125.

[7] S. Andrews, I. Tsochantaridis, T. Hofmann, Support Vector Machines for Multiple-Instance Learning, in: Thrun, K. Obermayer (Eds.), Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2003, pp. 561–568.

[8] A. Zafra, M. Pechenizkiy, S. Ventura, Hydr-mi: A hybrid algorithm to reduce dimensionality in multiple instance learning, Inf. Sci. 222 (2013) 282–301.

[9] A. Zafra, M. Pechenizkiy, S. Ventura, Relieff-mi: An extension of relieff to multiple instance learning, Neurocomput. 75 (1) (2012) 210–218.

[10] A. Ben-Hur, D. L. Brutlag, Remote homology detection: a motif based approach, in: ISMB (Supplement of Bioinformatics), 2003, pp. 26–33.

[11] R. Saidi, S. Aridhi, E. M. Nguifo, M. Maddouri, Feature extraction in protein sequences classification: a new stability measure, in: Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, BCB '12, ACM, New York, NY, USA, 2012, pp. 683–689.

[12] R. Saidi, M. Maddouri, E. M. Nguifo, Protein sequences classification by means of feature extraction with substitution matrices, BMC Bioinformatics 11 (2010) 175.

[13] T. G. Dietterich, R. H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artif. Intell. 89 (1-2) (1997) 31–71.

[14] K. Liolios, K. Mavromatis, N. Tavernarakis, N. C. Kyrpides, The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata, Nucleic Acids Research 36 (suppl 1) (2008) D475–D479.

[15] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.

[16] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool., Journal of molecular biology 215 (3) (1990) 403–410.

[17] H. Nishida, M. Nishiyama, Evolution of lysine biosynthesis in the phylum deinococcus-thermus., Int J Evol Biol 2012.

[18] N. Ohtani, M. Tomita, M. Itaya, An extreme thermophile, thermus thermophilus, is a polyploid bacterium., J Bacteriol 192 (20) (2010) 5499–505.

[19] N. Gao, B.-G. Ma, Y.-S. Zhang, Q. Song, L.-L. Chen, H.-Y. Zhang, Gene expression analysis of four radiation-resistant bacteria, Genomics Insights 2 (2009) 11–22.

ICEFinder: Knowledge-based Identification of Integrative Conjugative Elements (ICEs) in newly sequenced genomes using coding sequences as tags.

Nathalie LEBLOND-BOURGET[1,2], Gérard GUEDON[1,2], Sophie PAYOT[1,2], Yann AUBERT[1,2,3,4], Malika SMAÏL-TABBONE[3,4] and Marie-Dominique DEVIGNES[3,4]

[1] DynAMic, UMR1128 Université de Lorraine, [2] DynAMic, UMR1128 INRA, [3] LORIA, UMR7503 Université de Lorraine , [4] LORIA, UMR7503 CNRS, Campus Scientifique BP239, 54506, Vandœuvre-lès-Nancy, France

{bourget, guedon, payot}@nancy.inra.fr

{malika, devignes, yann.aubert}@loria.fr

Abstract *Integrative Conjugative Elements (ICEs) and Integrative Mobilisable Elements (IMEs) are increasingly recognized as important mediators of horizontal gene transfer among bacteria. We present here the first steps of a new method called ICEFinder for ICE and IME identification in Firmicutes, based on our knowledge of these elements within* Streptococcus. *In brief, the strategy consists in detecting in bacterial genomes a tri-partite tag composed of regions coding for a coupling protein, a relaxase and an integrase. This requires first to filter all coding sequences on the basis of the similarity of their products with coupling proteins, relaxases and integrases from known ICEs and IMEs using BLASTP. After validation, visualization of the tag topology on the genome facilitates expert authentication of ICEs or IMEs. This strategy has been applied to 72 sequenced streptococcal genomes and has led to the identification of possible ICEs or IMEs in 50 genomes. Moreover, it has extended our knowledge on domain architecture of coupling proteins, relaxases and integrases.*

Keywords *Integrative Conjugative Elements, genomic islands, Gram positive bacteria, firmicutes, sequence annotation, database, domain architecture*

1 Context

Horizontal transfer of mobile genetic elements plays a key role in bacterial evolution. Besides well-known elements, such as conjugative plasmids or prophages, more and more data suggest that elements belonging to two poorly known classes, the integrative and conjugative elements (ICEs) and the integrative and mobilizable elements (IMEs), are widespread [1, 2, 3]. These elements, integrated in bacterial chromosomes, can excise, form a circular intermediate, transfer to a recipient cell by conjugation and reintegrate into the chromosome. They have a modular structure including at least two functional units allowing their maintenance and dissemination (recombination and conjugation modules). The ICEs encode an integrase that catalyses its own excision, generally by site-specific recombination between short direct repeats flanking the ICE, and its own integration, generally in a specific site. The conjugation module includes an origin of transfer (*oriT*), a relaxase that nicks *oriT* to initiate the transfer and proteins required for the synthesis of the transmembrane conjugation pore. One of them, the coupling protein interacts with the DNA-relaxase complex and couples it with the conjugation pore. Like ICEs, IMEs are autonomous for their excision and integration. They also carry an incomplete conjugation module that, at least, includes an *oriT* and a relaxase gene. They need the conjugation pore encoded by an ICE or a conjugative plasmid to transfer (mobilization in *trans*).

The ICEs and IMEs were very rarely searched for themselves in genomes. They are difficult to detect as conjugation genes are generally poorly annotated. Furthermore, none of the ICE/IMEs modules is specific of ICEs/IMEs leading to confusion with integrated plasmids or prophages; only their combined presence is characteristic of these elements [4]. The aims of our study is to construct a method for specifically identifying ICEs and IMEs and evaluating the prevalence and diversity of ICEs and IMEs first in complete genomes of *Streptococcus* and then in other Firmicutes. The genus *Streptococcus* constitutes an interesting model to study genome evolution as they include bacteria used in dairy industry, commensals, and human and/or animal pathogens.

2 Methods

Sets of reference proteins were constituted manually as part of well established and distinct ICEs from various Firmicutes. The reference sequences were blasted against all protein products encoded by 72 sequenced genomes from the genus *Streptoccoccus* using KoriBlast. Hits were filtered according to identity percentage and query coverage. Domain composition was determined in batch using the standalone version of InterProScan-5. Hits were then validated manually by the expert. This procedure was repeated once to check for possible further enrichment in new sequences at the second BlastP round. Finally, putative ICE tags were identified: either monopartite (coupling protein, relaxase or integrase alone) or bipartite (a combination of two of these protein types) or tri-partite (all three types of proteins). The ICE database (ICEdb) was designed and built using Postgresql management system to store and organize all collected data at each stage of the procedure. The relative positions of ICE tags on the genome were visualized using Artemis. Protein sub-types were studied with respect to their domain architecture.

3 Results and Discussion

The diversity of sequences present in the protein datasets was estimated by the number of clusters obtained at 90% sequence identity and compared before and after the two Koriblast rounds. It revealed a ~2-fold increase for coupling proteins (from 28 to 58 clusters) and for relaxases (from 40 to 87 clusters), a ~3-fold increase for Tyr-integrases (from 42 to 122 clusters) and for Ser-integrases (from 12 to 41 clusters) and no increase at all for DDE-transposases (9 clusters before and after Koriblast enrichment).

We then studied domain architecture using InterProScan analysis of one representant of each cluster. The number of distinct architectures is quite limited (from 4 to 11 depending on protein type). One first result of our methodology is thus to identify distinct sub-types for each type of ICE/IME protein, some of which never described before, as for instance those involving Pfam domain PF01719 (IPR002631 : “Plasmid replication protein”), only found in IMEs.

The distribution of all CDSs coding for coupling proteins, relaxases and integrases across the 72 genomes considered in this study reveals new tri-partite tags in 40 genomes and new bi-partite tags in 10 genomes. In 19 other genomes only monopartite tags are found, mostly integrase likely pertaining from some prophages. One genome contains no ICE/IME tag at all. The two remaining genomes contain reference tags but are not enriched in additional ones by the procedure. In total our method allows further investigation of 50 *Streptococcus* genomes displaying novel tri- or bi-partite tags.

Visual inspection of ICE/IME tags was performed using Artemis, the well-known prokaryotic genome browser and annotation tool. Well-established types of ICE were recognized as well as interesting atypical constructs involving aggregated ICEs and IMEs. Attempts to delineate ICE/IME borders are underway, based on the search for direct repeats. Comparison with existing ICE or genomic islands resources reveals that our ICEFinder strategy is capable of discovering new ICEs and IMEs which get validated by the expert.

The ICEFinder strategy is based on the identification of ICE/IME tags in the *Streptococcus* genomes. It is a knowledge-based approach because it is based on (i) expert selection of initial sets of reference proteins and (ii) expert validation of BLASTP hits. From a computer-science point of view, the method described here is analogous to label propagation in semi-supervised learning. Indeed, the curated ICEdb resource built in this study represents an original opportunity for machine learning approaches to be set up for classifying ICEs and IMEs on the basis of domain composition, island topology and CDS content, integration site, etc. thus improving ICE and IME identification in genomes.

Acknowledgements

This work was supported by the Region Lorraine and French Government through the PRST MISN (Modelisation, Information et Systemes Numeriques), and by the “Structure Federative EFABA”.

References

[1] V. Burrus, G. Pavlovic, B. Decaris, G. Guedon, The ICESt1 element of *Streptococcus thermophilus* belongs to a large family of integrative and conjugative elements that exchange modules and change their specificity of integration. *Plasmid*, 48:77-97, 2002.

[2] M. Brochet, E. Couve, P. Glaser, G. Guedon, S. Payot, Integrative conjugative elements and related elements are major contributors to the genome diversity of *Streptococcus agalactiae*. *J. Bacteriol.*, 190:6913-7, 2008.

[3] J. Guglielmini, L. Quintais, M.P. Garcillan-Barcia, F. de la Cruz, E.P. Rocha, The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* 7:e1002222, 2011.

[4] V. Burrus, G. Pavlovic, B. Decaris, G. Guedon, Conjugative transposons: the tip of the iceberg. *Mol. Microbiol.*, 46:601-10, 2002.

In silico definition of new lignin peroxidase classes in fungi and putative relation to pathogeny

Nizar Fawal[1,2], Christophe Roux, Catherine Mathé[1,2], Christophe Dunand[1,2]

[1]Université de Toulouse, UPS, UMR 5546, Laboratoire de Recherche en Sciences Végétales, BP 42617, F-31326 Castanet-Tolosan, France.

[2]CNRS, UMR 5546, BP 42617, F-31326 Castanet-Tolosan, France.

{fawal, roux, mathe, dunand }@lrsv.ups-tlse.fr

Abstract *The explosion of genomic projects from fungi allows performing exhaustive and expert annotation of ligninase encoding sequences. Large numbers of annotated sequences not belonging to the well described basidiomycete ligninases (LiP, MnP and VP) have been also found in Ascomycota but not in basal fungi. In addition to the LiP, MnP and VP classes, six new classes have been detected: three found in plant pathogen Ascomycota and three in Basidiomycota. Class II peroxidases (CII) from Ascomycetes are rarely subjected to duplications unlike those from Basidiomycetes which can form large duplicated families. Even if these CII form two well distinct clusters with divergent gene structures, they share the same conserved key residues suggesting that they evolved independently from similar ancestral sequences with few or no introns. The lack of ligninase sequences in basal fungi, some Ascomycota and some Basidiomycota, together with the absence of duplicated CcP in fungi containing ligninases, suggests the potential emergence of an ancestral ligninase sequence from the duplicated CcP after the separation between Ascomycota and Basidiomycota.*

Keywords Evolution, Genome Analysis, Comparative genomics.

1 Introduction

Plant cell walls represent the major components of plant biomass and play important roles in plant biology, environment, and various industries. Cellulose, hemicelluloses, pectins, proteins and lignins are the main components of cell walls. Altogether, they form intricate macromolecular networks that are synthesized through complex processes. Lignocellulosic cell walls are energetic source for human industries but also for several other micro-organisms such as fungi that use these cell walls as a source of carbon. They are decayed primarily by wood and litter decomposers that easily digest cellulosic compounds thanks to enzymatic activities. However, the accessibility to these compounds is restricted by the presence of more resistant compounds, lignin polymers. Therefore, efficient lignin depolymerization capacities are necessary and have been developed by wood-decaying basidiomycetes such as white-rot fungi. These fungi can either degrade both lignin and polysaccharides (simultaneous decay) or preferentially remove lignin, leaving most of the cell wall polysaccharides unaffected (selective decay). For example, *Phanerochaete chrysosporium* or *Coprinopsis cinerea* are very efficient wood decomposers: they can simultaneously degrade lignin and cellulose [1,2] . A closely related species, *Ceriporiopsis subvermispora*, also depolymerizes lignin with reduced cellulose degradation [1]. Several other white-rot fungi genomes have been screened for the presence of ligninolytic peroxidases such as *Agaricus bisporus* or *Pleurotus ostreatus* which possess limited copies of ligninases [3].

Postia placenta and *Serpula lacrymans*, two brown-rots fungi, mainly capable of cellulose break-down, have only few or no ligninase, also known as class II peroxidases [4]. The lack of this class is also observed in the ectomycorrhizal *Laccaria bicolor* [5]. Therefore, the reduced number or absence of peroxidase isoforms could be correlated with a mode of acquisition of carbonaceous materials.

Phylogenetic analysis of lignin degrading peroxidases from agaromycetes and from Hymenochaetales [6] suggest that lignin degradation by white rot fungi is strongly correlated with the presence of ligninolytic peroxidases and could not be performed by laccases only. The ligninases, or class II peroxidases, are extracellular fungal peroxidases, initialy subdivided into 3 groups: lignin (LiP), manganese (MnP) and versatile (VP) peroxidases [7] and belong to the superfamily of non-animal peroxidase together with the class I and III peroxidases. In addition, recent studies have defined a fourth group found in basidiomycetes so called "generic peroxidases" (GP). The analysis of numerous available basidiomycete and ascomycete genomes pinpointed that ligninase classification needed to be updated. Recent comparative analyses performed on 31 basidiomycete genomes have confirmed the complexity of the class II peroxidases evolution and classification [8]. Interestingly, these studies did not include putative ascomycete ligninases. Besides, some orders appeared to be under-represented with only one organism while others are over-represented with 15 organisms. These heterogeneous distributions of the class II sequences among orders can introduce a bias regarding the conclusion about the sequences' gains and losses. Based on the hypothesis of common ancestral class I/II/III peroxidases and to complete this work, we decided to analyse all basidiomycete and ascomycete genomes publicly available from NCBI and JGI. We also performed manual and expert annotation to get rid of automatic annotation errors.

Until now, Class II peroxidases mainly described in basidiomycetes were considered as tools for lignin degradation to provide carbon to fungi. The presence of low copy ligninases like in ascomycetes opened a new field of investigations regarding a new putative function for the class II peroxidases. Finally, Class II peroxidases could be also a means to invade host cells.

2 Material and method

2.1. Data mining and annotation with scipio

An exhaustive data mining procedure has been done on the Fungi genomes available in JGI and NCBI to extract the ligninase sequences. All sequences used in this study have been annotated by an expert process to discard prediction errors subsequent to automatic annotations. First, a comparison between all predicted proteins from JGI and the genomic sequences (NCBI) is done with the tblastn program from the NCBI BLAST package. This allows the creation of initial batches corresponding to the different protein families. Then, a manual and individual curration of each predicted sequence is performed: gene structure, length and the presence of key residues are mandatory controls. EST libraries, if available, have also been used to confirm annotations. These protein batches were then used to precisely determine the corresponding chromosome positions, gene structures and CDS sequences with a protocol based on Scipio [9]. New paralogs, not initially annotated, were found with the help of this procedure.

2.2. Phylogenetic and clustering analysis

All protein sequences used for the *in silico* analyses are available from the PeroxiBase database (http://peroxibase.toulouse.inra.fr) [10,11]. First, protein sequences were aligned using MAFFT. Then, these alignments were furthermore inspected and visually adjusted with BioEdit [12].

Then, phylogenies were estimated by maximum-likelihood (ML) using PhyML-aLRT [13] The substitution model determined by protTest [14] was WAG [15] which takes into account the evolutionary relationships within each family. The maximum-likelihood algorithm BIONJ [16] distance-based tree was used to refine the starting tree. The latter was optimized for topology, branch lengths and rate parameters. The aLRT statistics were used to check whether the branch being studied provided a significant gain in comparison with the null hypothesis that involves collapsing that branch but leaving unchanged the rest of the tree topology [17]. The aLRT statistics were interpreted using the non-parametric Shimodaira-Hasegawa-like procedure.

Finally, trees were edited and analysed using Archeopterix [18]. The annotation files needed for the tree annotation were produced from the multi-alignment files.

2.3. Gene structure analysis

The intron/exon coordinates together with the corresponding genomic sequences of all identified genes were determined with Scipio [9]. The conserved intron/exon organization of the different families was verified with software that position introns on protein sequences: CIWOG [19] and GECA [11] analyse the evolution or conservation of introns between paralogs as well as between species.

Intron size changes were visualized through a graphical representation provided by GECA [11].

2.4. Conserved Common Cintrons analysis

Conserved common introns from all Basidiomycetes sequences with gene structure were analysed. First, they were aligned altogether with MAFFT to produce a protein alignment which was completed with the identification of common introns (or cintrons) in the corresponding genes with CIWOG. Cintrons were extracted from CIWOG's database and only those present in one or several sub-classes with a conservation rate higher than 40% were considered conserved. Finally, the sequences were placed in order of appearance in the phylogenetic tree and the conserved cintrons were highlighted for each sequence.

2.5. New PROSITE profiles design and WebLogo

Based on a global phylogenetic analysis, different protein clusters have been defined to update the existing PROSITE profiles [10] and to design new specific profiles using a strategy based on the silencing of residues. These profiles were built from full length alignment of each protein cluster. First, all the sequences from the different protein clusters were merged together to construct general alignments. The alignments were then simply split without modifying the alignment of residues into several sub-alignments according to their cluster definition. Each cluster alignment would contain an annotation line where residues conserved in the whole family are tagged. This annotation line is then used by the profile construction program to down weigh family-conserved columns therefore only cluster specific residues are taken into consideration. The reliability of the cluster is feature supported by the analysis of the gene structure together with the presence or absence of the key residues specific to the well described LiP, MnP and VP families. Furthermore, the sets of cluster proteins have been used to create graphical sequence logos corresponding to amino acids frequency using Weblogo3 [20]. 300 Class II peroxidases from Basidiomycota and Ascomycota were aligned with MAFFT, and then separated into groups according to their classification. Weblogos were created for each group with Weblogo3 and aligned manually with the others in order to easily identify the conserved amino acids between the subfamilies.

3. Results and discussion

3.1. New classes of ligninases

Automatic genome annotation has been demonstrated to produce misprediction in the case of multigenic families and genes containing numerous exons and introns that are sometimes very short. The class II peroxidases familie can be large and contains up to 15 introns in a sequence, with short exons and introns (e.g: 6 nt for the last exon) . The quality of the annotation is necessary to perform a global evolutive analysis of mutigenic families such as those of the class II peroxidases. 134 genomes from ascomycetes, 54 genomes from basidomycetes and 8 genomes from more basal fungi have been extensively annotated for lignases encoding sequences which allowed identifying 391 sequences used for the phylogenetic and clustering analysis. Marginal ligninase sequences, obtained from organisms without genome sequencing projects, have also been included in the phylogenetic and gene structure analysis and helped to support isolated phylogenetic position of some organisms. Characteristic residues necessary for heam binding and electron transport are detected in all class II peroxidases analysed (Fig. 1). However, sequences detected in ascomycetes formed a group distinct from the well described basidiomycete ligninases (LiP, MnP and VP). Some species such as *Alternaria brassicicola* or *Colletotrichum graminicola* contained up to three different copies of ligninases. The clear distribution in these three clusters (Fig. 2) helps to define and design three sub-classes of ascomycete ligninases, thereafter referred as AA, AB and AC, and their corresponding profiles (Fig. 3). Members of these new classes are only detected in Pezizomycotina and are absent from the other ascomycetes sub-phylums. Moreover, in the Pezizomycotina sub-phylum, ligninase encoding sequences are not detected in all species. They can be found in species known to interact with plants, in a pathogenic manner as well as in a saprophytic manner (Table 1). The ascomycetes class II is monophyletic and is well separated from the large basidiomycetes class II cluster.

The situation for the basidiomycetes is much more complex. Three classes have been largely described based on their *in vitro* catalyzing activities namely the lignin peroxidases (LiP), the manganese peroxidases (MnP) and the versatiles peroxidases (VP). Recently a fourth group so called "generic peroxidases" (GP) has been detected [22]. The exhaustive genome mining of 53 basidiomycete genomes has demonstrated the need to redefine the existing profiles. Furthermore, it has put in evidence sequences belonging neither to the three classes previously mentioned nor to the new fourth group. Therefore, within this fourth group, three new basidiomycete ligninase sub-classes have been defined, latter called BA, BB and BC.

In basidiomycetes, one can observe extensive gene duplication for LiP, MnP and the three other new classes. All these duplications are very recent since they form well-supported clusters specific for each specie. It begs the question of duplication events widespread among basidiomycetes but it appears that other peroxidase classes such Cytochrome C peroxidases (CcP) or glutathione peroxidases were not subjected to duplication. This suggests that these duplications are an evolutionary response to selection pressure.

Also, by comparing the conserved residues found in Basidiomycota and Ascomycota (Weblogo), we can see that (i) there are many conserved residues (~23), dispersed throughout the sequences, between these two phyla. (ii) The 3 sub-classes defined in Ascomycota do not share any of the defined residues with catalytic properties found in Basidiomycota. (iii) The Mn^{2+}-binding site formed by Glu36, Glu40 and Asp175 responsible for the enzymatic oxidation of Mn^{2+} to Mn^{3+} thought to be a unique characteristic of MnP and VP is also found in the 3 new sub-classes defined in Basidiomycota. (iv) The new sub-class BA of Basidiomycota should be able to oxidize high redox-potential aromatic compound via longue range electron transfer, as LiP and VP do, since Trp171 is also present in this sub-class.

No ligninase like sequence has been detected in several basal fungi analysed. Nevertheless, key residues described in MnP, LiP, and VP are detected in ascomycetes sequences suggesting a common ancestral sequence. This suggests that class II peroxidases detected in ascomycetes and basidiomycetes appeared later in the fungal lineage and are coming from a common ancestor.

In order to support the phylogenetic analysis, gene structures of the different sub-classes of Basidiomycota have been compared. The intron positions belonging to 41 organisms, were extracted from CIWOG's database and aligned in their order of appearance in the phylogenetic tree. Out of 62 cintrons detected, 29 were considered conserved since they were present in one or several sub-classes with a conservation rate higher than 40% (Data not shown). The analysis clearly reveals that LiP, VP and BB which are close to each other, share a similar intron/exon structure. On the other hand, MnP, BC and BA have each a unique gene structure.

3.2. Working hypothesis of ligninase evolution and different purpose for ligninases

Ligninases are detected in Basidiomycota and Ascomycota but are absent from basal fungi. Since Basidiomycota and Ascomycota are monophyletic sister groups which emerged from a common ancestral organism, ligninase from these two groups evolved independently from a same ancestral sequence (convergent evolution). Besides, ligninases are class II peroxidases, thus belonging to the same superfamily, and sharing common key residues , with class I Cytochrome C peroxidases (CcP) [23]. When comparing Basidimycota and Ascomycota for the presence of CcP and ligninases (Table 1), we can observe that: (i) in most of the cases, two CcP sequences can be detected in fungi which do not contain ligninase sequenceand (ii) only one CcP is detected when the genome contains at least one ligninase sequence (Table 1). This suggests that the ancestral sequence could be a CcP. A similar theory of evolution has already been described for the class III peroxidases [24]. Indeed, class III peroxidases were only detected in plants which lack CcP. Class III and ligninases are both subjected to numerous species specific duplications (tandem and segmental, data not shown), contain highly conserved Cysteine residues necessary for disulfite bridges and are detected in specific species. A second theory can be discussed here, it is based on the idea that all fungi possessed at least one class II sequence and that the loss event occurred more recently. On the principle of maximum parsimony, this theory would require many discrete events of gene loss, seems unlikely.

Even if numerous key residues are well conserved between ligninases from all fungi, residues described as necessary for electrons transfert are missing and the position of the conserved Cysteines varies between ascomycete and basidiomycete ligninases. These divergences lead to suggest divergent functions (or different catalytic mechanism). Lignivor fungi possess large batteries of ligninases encoding sequences which enables them to cut the lignin and then to use it as a source of carbone. Saprophyte fungi, described as not being lignivor, present low numbers of ligninase sequences. They probably use them just to cut the lignin in order to increase the accessibility to other cell wall components such as cellulose and hemicellulose. Finally, plant pathogens ascomycetes, which also contain from 1 to 7 ligninase encoding sequences, use these proteins to cut the lignin in order to infect the host cell. Whithin saprophytes, another dichotomy can be observed between biotroph and necrotroph: the first present only few class II copies whereas the second contain up to 7 copies.

Furthermore, the intron positions and numbers are not conserved between ascomycetes and basidiomycetes (Data not shown). Basidiomycetes contains more introns than ascomycetes (on average 8 and 2 respectively), but their sizes on average are higher in ascomycetes (74 nt) than in basidiomycetes (54 nt). Similar phenomena is also observed with other families such as glutathion peroxidases (1-2 for asco and 3 to 6 for basidio). Altogether, this suggests that two ancestral sequences should contain no or few introns. Marginal intron gain and loss are detected when the families are subjected to numerous duplication events.

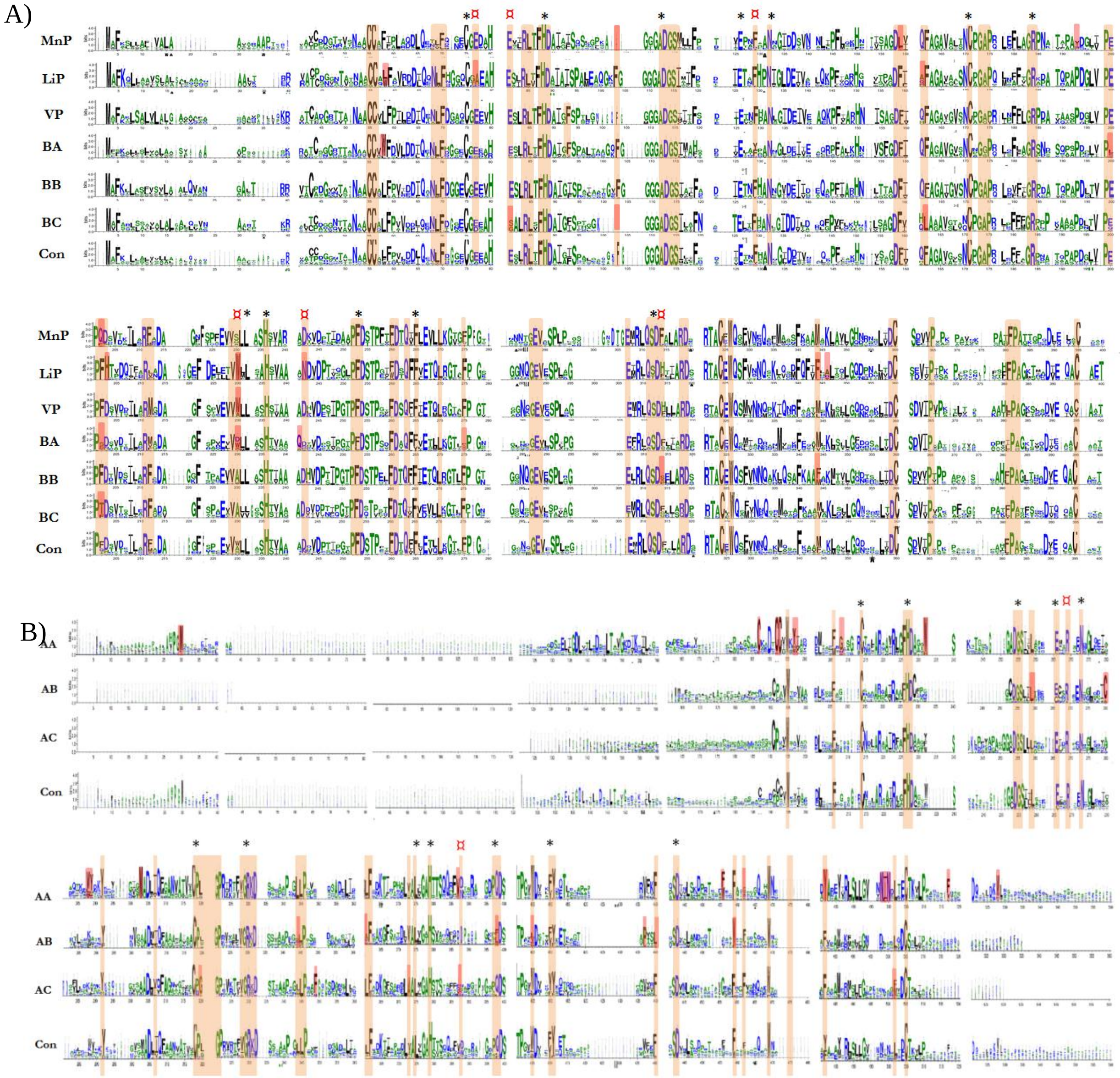


Fig 1. Weblogo of the different Basidiomycota (A) and Ascomycota (B) class II peroxidase classes. 300 Basidiomycota and Ascomycota Class II peroxidases were aligned with Mafft, then separated into subfamily groups. Weblogos were created for each group with Weblogo3 and aligned manually with the others in order to easily identify the conserved Amino Acids between the subfamilies which are highlighted in orange and those that are specific to one or several subfamilies highlighted in red. AA: Ascomycete class II peroxidases type A, AB: Ascomycete class II peroxidases type A, AC: Ascomycete class II peroxidases type B, BA: Basidiomycete class II peroxidases type C, BB: Basidiomycete class II peroxidases type A, BC: Basidiomycete class II peroxidases type C, VP: Versatile peroxidases, MnP: Manganese peroxidases, LiP: Lignin peroxidases. * Conserved residues between Basidiomycota and Ascomycota class II peroxidase classes.

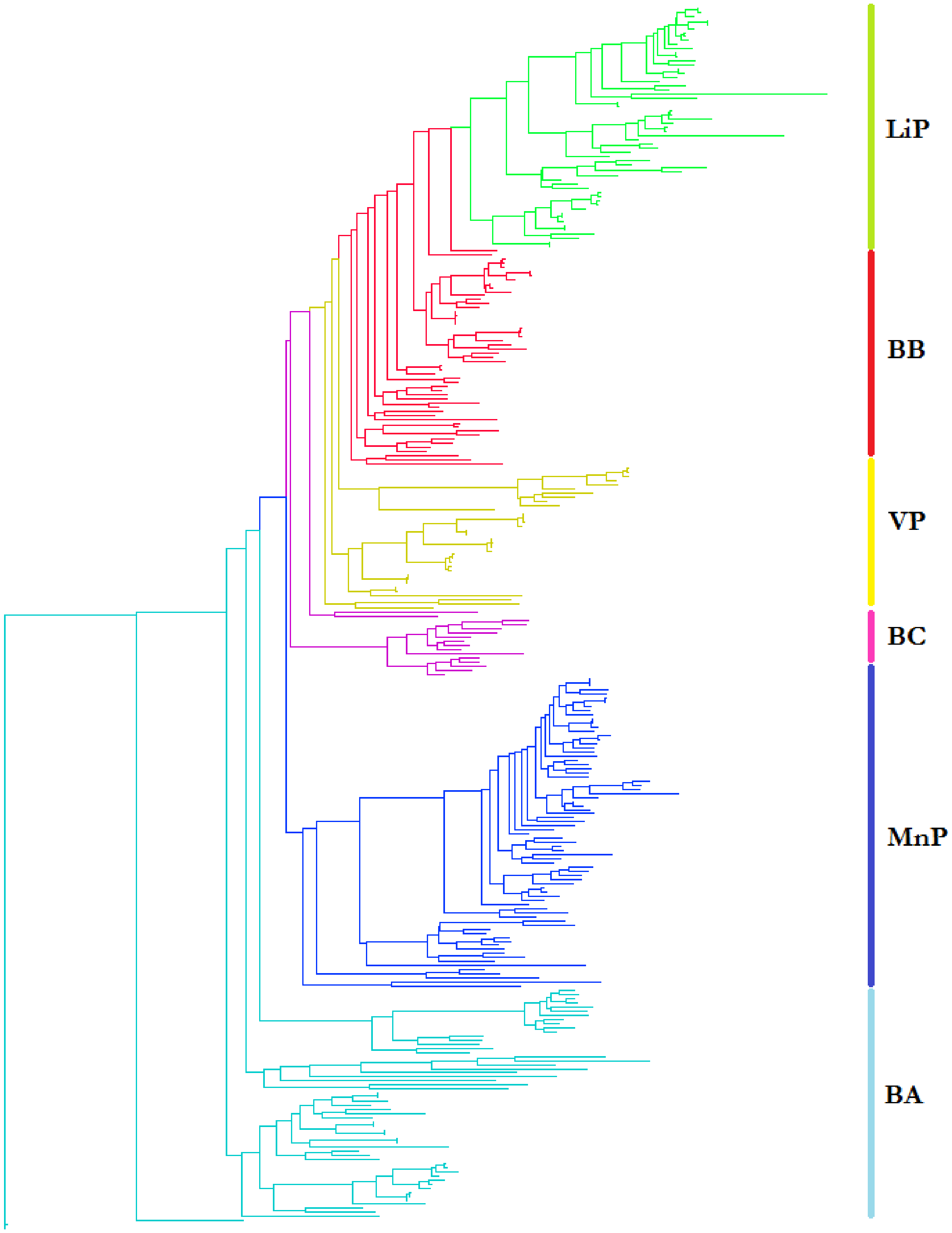


Fig.2. Phylogenetic representation of the class II peroxidases from Basidiomycota.

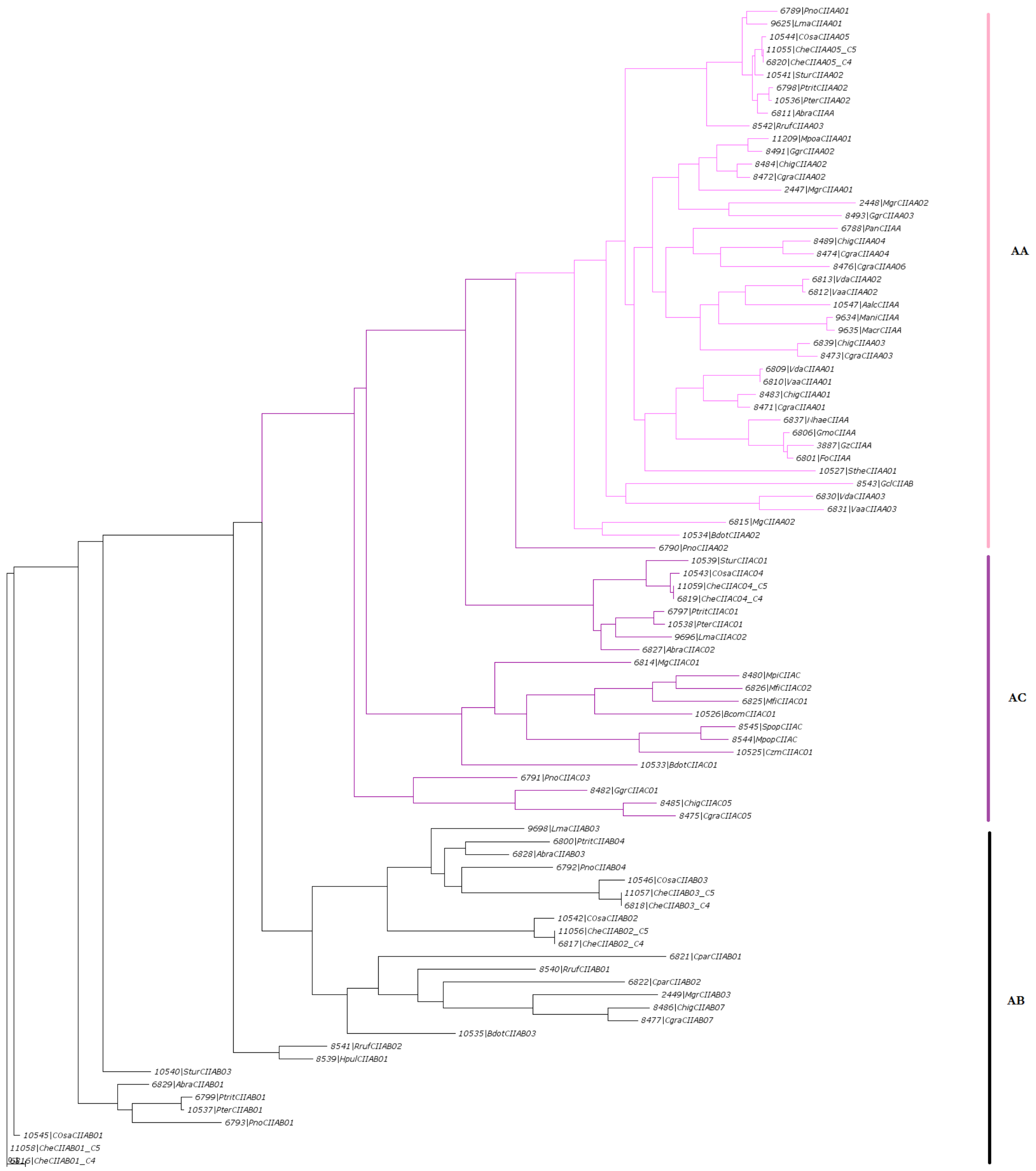


Fig. 3. Phylogenetic representation of the class II peroxidases from Ascomycota.

	Phylum	Class		Presence CII	Presence CcP and #
Basidiomycota					
	Agaricomycotina	Agaricomycetes	Agaricales	(+)	1
			Auriculariales	(+)	1
			Boletales	(-)	1
			Cantharellales	(-)	1
			Corticiales	(+)	1
			Gloeophyllales	(-)	1
			Hymenochaetales	(+)	1
			Polyporales	(+)	1
			Russulales	(+)	1
		Tremellomycetes		(-)	2
	nd	Wallemiomycetes		(-)	1
	Pucciniomycotina	Microbotryomycetes		(-)	2
		Mixiomycetes		(-)	2
		Pucciniomycetes		(-)	2
	Ustilaginomycotina	Exobasidiomycetes		(-)	2
		Ustilaginomycetes		(-)	2
Ascomycota					
	Pezizomycotina				
		Dothideomycetes		(+)	1
		Eurotiomycetes		(-)	2
		Lecanoromycetes		(-)	1
		Leotiomycetes		(-)	2
		Sordariomycetes		(+)	1
	Saccharomycotina			(-)	2 or 1
	Schizosaccharomycotina			(-)	0
	Taphrinomycotina			(-)	2

Table 1. Presence of Cytochrome C peroxidases (CcP) and class II peroxidases in Ascomycota and Basidiomycota. The presence of class II peroxidase is shown by a "+" and it's absence by a "-".

References

[1] Fernandez-Fueyo E, Ruiz-Dueñas FJ, Ferreira P, Floudas D, Hibbett DS, Canessa P, Larrondo LF, James TY, Seelenfreund D, Lobos S, Polanco R, Tello M, Honda Y, Watanabe T, Ryu JS, San RJ, Kubicek CP, Schmoll M, Gaskell J, Hammel KE, St John FJ, Vanden Wymelenberg A, Sabat G, Splinter BonDurant S, Syed K, Yadav JS, Doddapaneni H, Subramanian V, Lavín JL, Oguiza JA, Perez G, Pisabarro AG, Ramirez L, Santoyo F, Master E, Coutinho PM, Henrissat B, Lombard V, Magnuson JK, Kües U, Hori C, Igarashi K, Samejima M, Held BW, Barry KW, LaButti KM, Lapidus A, Lindquist EA, Lucas SM, Riley R, Salamov AA, Hoffmeister D, Schwenk D, Hadar Y, Yarden O, de Vries RP, Wiebenga A, Stenlid J, Eastwood D, Grigoriev IV, Berka RM, Blanchette RA, Kersten P, Martinez AT, Vicuna R, Cullen D (2012) Comparative genomics of Ceriporiopsis subvermispora and Phanerochaete chrysosporium provide insight into selective ligninolysis. Proc Natl Acad Sci U S A 109: 5458-5463.

[2] Stajich JE, Wilke SK, Ahrén D, Au CH, Birren BW, Borodovsky M, Burns C, Canbäck B, Casselton LA, Cheng CK, Deng J, Dietrich FS, Fargo DC, Farman ML, Gathman AC, Goldberg J, Guigó R, Hoegger PJ, Hooker JB, Huggins A, James TY, Kamada T, Kilaru S, Kodira C, Kües U, Kupfer D, Kwan HS, Lomsadze A, Li W, Lilly WW, Ma LJ, Mackey AJ, Manning G, Martin F, Muraguchi H, Natvig DO, Palmerini H, Ramesh MA, Rehmeyer CJ, Roe BA, Shenoy N, Stanke M, Ter-Hovhannisyan V, Tunlid A, Velagapudi R, Vision TJ, Zeng Q, Zolan ME, Pukkila PJ. (2010) Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom Coprinopsis cinerea (Coprinus cinereus). Proc Natl Acad Sci U S A. 29: 11889-94.

[3] Ruiz-Dueñas FJ, Fernández E, Martínez MJ, Martínez AT. (2011) Pleurotus ostreatus heme peroxidases: an in silico analysis from the genome sequence to the enzyme molecular structure. C R Biol. 11:795-805.

[4] Martinez AT (2002) Molecular biology and structure function of lignin-degrading heme peroxidase. Enz Microb Technol 30: 425-444

[5] Martin F, Selosse MA. (2008) The Laccaria genome: a symbiont blueprint decoded. New Phytol. 180(2):296-310.

[6] Morgenstern I, Klopman S, Hibbett DS (2008) Molecular evolution and diversity of lignin degrading heme peroxidases in the Agaricomycetes. J Mol Evol 66: 243-257

[7] Welinder KG (1992) Plant peroxidases: structure-function relationships. In: Penel C, Gaspar T, Greppin H (eds) Plant Peroxidases. University of Geneva, Switzerland, pp 1-24

[8] Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martínez AT, Otillar R, Spatafora JW, Yadav JS, Aerts A, Benoit I, Boyd A, Carlson A, Copeland A, Coutinho PM, de Vries RP, Ferreira P, Findley K, Foster B, Gaskell J, Glotzer D, Górecki P, Heitman J, Hesse C, Hori C, Igarashi K, Jurgens JA, Kallen N, Kersten P, Kohler A, Kües U, Kumar TK, Kuo A, LaButti K, Larrondo LF, Lindquist E, Ling A, Lombard V, Lucas S, Lundell T, Martin R, McLaughlin DJ, Morgenstern I, Morin E, Murat C, Nagy LG, Nolan M, Ohm RA, Patyshakuliyeva A, Rokas A, Ruiz-Dueñas FJ, Sabat G, Salamov A, Samejima M, Schmutz J, Slot JC, St John F, Stenlid J, Sun H, Sun S, Syed K, Tsang A, Wiebenga A, Young D, Pisabarro A, Eastwood DC, Martin F, Cullen D, Grigoriev IV, Hibbett DS (2012) The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science 336: 1715-1719

[9] Keller O, Odronitz F, Stanke M, Kollmar M, Waack S (2008) Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics 9: 278.

[10] Koua D, Cerutti L, Falquet L, Sigrist CJA, Theiler G, Hulo N, Dunand C (2009) PeroxiBase: a database with new tools for peroxidase family classification. Nucleic Acids Research 37: D261-D266

[11] Fawal N, Savelli B, Dunand C, Mathé C (2012) GECA: a fast tool for gene evolution and conservation analysis in eukaryotic protein families. Bioinformatics 28: 1398-1399

[12] Tippmann HF. (2004) Analysis for free: comparing programs for sequence analysis. Brief Bioinform. 5(1):82-7.

[13] Guindon S, Gascuel O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52(5):696-704.

[14] Abascal F, Zardoya R, Posada D. (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 21(9):2104-5.

[15] Whelan S, Goldman N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18(5):691-9.

[16] Gascuel O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 14(7):685-95.

[17] Anisimova M, Gascuel O. (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol. 55(4):539-52.

[18] Archaeopteryx **:** https://sites.google.com/site/cmzmasek/home/software/archaeopteryx

[19] Wilkerson MD, Ru YB, Brendel VP (2009) Common introns within orthologous genes: software and application to plants. Briefings in Bioinformatics 10: 631-644

[20] Crooks GE, Hon G, Chandonia JM, Brenner SE. (2004) WebLogo: a sequence logo generator. Genome Res. 14(6):1188-90.

[21] Martínez AT, Ruiz-Dueñas FJ, Martínez MJ, Del Río JC, Gutiérrez A. (2009) Enzymatic delignification of plant cell wall: from nature to mill. Curr Opin Biotechnol. 20(3):348-57.

[22] Larrondo L, Gonzalez A, Perez Acle T, Cullen D, Vicuña R. (2005) The nop gene from Phanerochaete chrysosporium encodes a peroxidase with novel structural features. Biophys Chem. 116(2):167-73.

[23] Welinder KG, Mauro JM, Nørskov-Lauritsen L. (1992) Structure of plant and fungal peroxidases. Biochem Soc Trans.;20(2):337-40.

[24] Passardi F, Zamocky M, Favet J, Jakopitsch C, Penel C, Obinger C, Dunand C. (2007) Phylogenetic distribution of catalase-peroxidases: are there patches of order in chaos? Gene; 397(1-2):101-13.

Gene history inferred from protein structures

The intricate evolution of GBD superfamily

Marc BERGDOLL[1]

[1] Service de biologie structurale de l'IBMP du CNRS, UPR2357, 12 rue du Général Zimmer,
67084, STRASBOURG, Cedex, France
marc.bergdoll@ibmp-cnrs.unistra.fr

Abstract *As evidenced already in 1986 by C.Chothia and A. Lesk, protein structures and particularly protein folds are better conserved througout evolution than their sequences. This property can be used to establish phylogenies based on structures superpositions more precisely than phylogenies based only on sequence alignments. The accumulation of known protein structures allows us today to enrich the concept of « structure-function relationship » by a new chapter « structure-phylogeny relationship ». We present here the case of a family where members sequence similarities are below the detection level but which has a unique ancestor and which is built upon the repetion, duplication/fusion and swap of a module about sixty amino-acids long.*

Keywords Structure based phylogeny, duplication/fusion of genes, domain swap, superfamily.

Histoire d'un gène inférée à partir de structures protéiques

L'évolution complexe de la superfamille GBD

Résumé *Comme C. Chothia et A. Lesk l'avaient mis en évidence, dès 1986, les structures des protéines et plus particulièrement leurs repliements sont mieux conservées au cours de l'évolution que leurs séquences. Cette propriété peut être mise à profit pour établir des philogénies structurales bien plus précises que des phylogénies basées uniquement sur des alignements de séquences. L'accumulation de structures protéiques résolues nous permet aujourd'hui d'enrichir le concept de « relation structure-fonction » et de lui adjoindre un volet « relation structure-phylogenie ». Nous présentons ici le cas d'une famille pour laquelle les similarités de séquences sont inférieures au seuil de détectabilité mais dérivant pourtant d'un unique ancêtre et batie sur la répétion, duplication/fusion et réorganisation d'un module d'environ soixante acides aminés.*

Mots-clés Phylogénie structurale, duplication/fusion de gènes, permutation de domaines, superfamille.

1 Introduction

The primary aim of protein structure determination is to understand their function illustrated by the well known concept of « structure-function relationship ». However the accumulation of solved structures deposited at the Protein Data Bank [1] allows new use of structures. More precisely, it enables the emergence of a new approach of phylogeny : structural phylogeny. Whenever applicable (i.e. when enough structures are available), this structure-based phylogeny prooves more powerful than phylogeny based solely on sequence alignments. This is no real surprise as structures are better conserved througout evolution than sequences [2]. Here I present the evolution of a family whose existence was discovered and history established only after the structure determination by X-ray crystallography of three proteins with unrelated functions and with sequence similarity far below the detection threshold. This family was later enriched by the addition of tens of relatives as structures of new proteins were solved, revealing unsuspected connections to the initial three-membered family. This example shows a growing potential of structures and nicely illustrates the concept of « structure-phylogeny relationship ».

2 The intricate history of VOC family

2.1 3D based identification of a new family

Some years ago the structure determination by X-ray crystallography of three different proteins with three different functions from three different organisms revealed surprisingly that they belong in fact to a same family [3]. Indeed, despite very low sequence similarity, undetected prior to structure resolution and superposition, Bleomycin Resistance Protein from *Streptoalloteichus hindustanus*, Dihydrogenase from *Burkholderia cepacia* and human Glyoxalase I share the same fold and core organisation. The biologicaly active entity is composed by the fourfold repetition of an elementary βαβββ domain. Structural comparisons, as well as phylogenetic analyses, strongly indicate that the modern family of proteins represented by these structures arose through a rich evolutionary history that includes multiple gene duplication and fusion events.

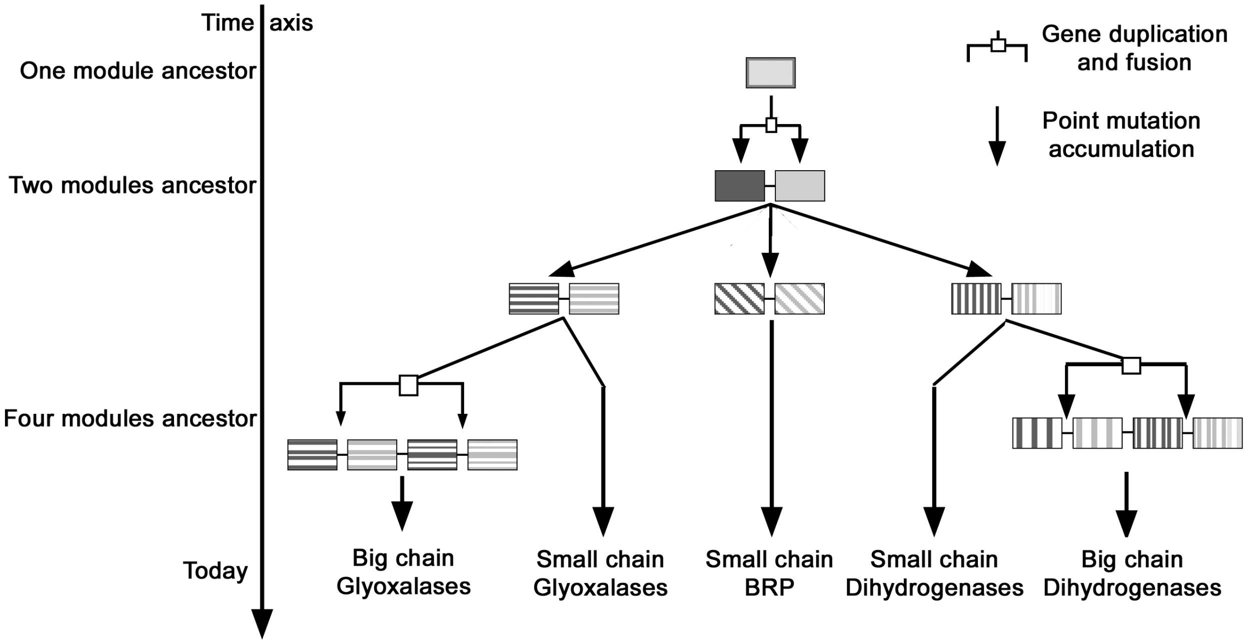


Figure 1. History of GBD superfamily.

2.2 From a small family to a big herd

Since then, about 80 new structures have been solved and submitted to the Protein Data Bank enriching the initial three membered GBD superfamily[1] by more than ten new families. New functions have been added as C3-degrading enzyme, toxoflavin-degrading enzyme, bleomycin acetyletransferase, lactoylglutatione lyase, TIOX, fosfomycin resistance protein, a virulence protein, methylmalonyl-CoA epimerase, NAD-dependent benzaldehyde dehydrogenase, mitomycin-binding protein, PHNB protein and several uncharacterized proteins. New organisms are now represented. Variations in the spatial arrangements of homologous modules are observed. New structural modules can be seen but are always associated to the initially identified βαβββ domain.

2.3 Actual family album

Here, I propose a new visit of the whole family using structures as well as sequences [4]. An evolutionary history is proposed that includes multiple gene duplication and fusion events as well as domain reorganisation and function acquisition or loss. These events appear to be historically shared in some cases, but parallel and historically independent in others. A significant early event is proposed to be the

[1] GBD family name is given as abbreviation of the Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase superfamily and fold, names introduced in SCOP database

establishment of metal-binding in an oligomeric ancestor prior to the first gene fusion. Variations in the spatial arrangements of homologous modules are described that are consistent with the structural principles of three-dimensional domain swapping, but in the unusual context of the formation of larger monomers from smaller dimers or tetramers. The comparisons support a general mechanism for metalloprotein evolution that exploits the symmetry of a homo-oligomeric protein to originate a metal binding site and relies upon the relaxation of symmetry as specific functions emerge. This work gives also some interesting insights into the general mechanism where constraints applied on homo-multimeric proteins are released by gene duplication/fusion and protein chain modifications (insertions, deletions, decorations) allowing emergence of new functionnalities.

3 Material and methods

Structures similar to the ones used to establish the initial family were searched for throughout the PDB with the Dali server [5]. Different templates were used for the search ranging from isolated βαβββ domain, single monomers (1BYL & 1FRO) composed of two βαβββ domains to the biological active units with four βαβββ domains (1BYL_dimer & 1HAN). The hits were then filtered to remove problematic or redundant structures (identical sequences or very similar sequences). 80 structures were finally kept with a total of 208 unique βαβββ domains (56 short chains with two domains and 24 long chains with four domains).

Individual domains were then extracted from these structures. Each domain was superimposed on all the other ones with PyMOL [6] using personnal scripts; RMSD and number of Cα were stored in a 208x208 matrix with RMSD between pairs in the triangle above the diagonal and number of superimposed Cα in the lower triangle). This matrix was then transformed by dividing each RMSD by the corresponding number of Cα into a normalized distance matrix used in R [7] to cluster the domains and draw the trees. The structural clustering was performed using the NJ method in APE library [8].

References

[1] PDB www.rcsb.org ; H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.

[2] C. Chothia, A.M. Lesk, The relation between the divergence of sequence and structure in proteins. *EMBO J.* Apr;5(4):823-6, 1986.

[3] M. Bergdoll, L. D. Eltis, A. D. Cameron, P. Dumas and J. T. Bolin, All in the family: Structural and evolutionary relationships among three modular proteins with diverse functions and variable assembly *Protein Sci.,* 7: 1661-1670, 1998

[4] Inferences about gene history from protein structural data: the complex evolution of GBD superfamily, M. Bergdoll (in preparation)

[5] L. Holm, P. Rosenström (2010) Dali server: conservation mapping in 3D. Nucl. Acids Res. 38, W545-549.

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

[7] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299-314, 1996.

[8] E. Paradis, J. Claude & K. Strimmer 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20:** 289–290. doi:10.1093/bioinformatics/btg412.

Exome Sequencing in Chronic Myelomonocytic Leukemia

Jane MERLEVÈDE[1], Nathalie DROIN[1], Margot MORABITO[1], Yannis DUFFOURD[2], Serge KOSCIELNY[3] and Éric SOLARY[1]

[1] Hématopoïèse normale et pathologique, UMR1009 INSERM, 114 rue Édouard Vaillant, Institut Gustave Roussy, 94805 Villejuif, France
{jane.merlevede, eric.solary}@igr.fr
[2] Plateforme de Génomique, 114 rue Édouard Vaillant, Institut Gustave Roussy, 94805 Villejuif, France
[3] Département de Biostatistiques, 114 rue Édouard Vaillant, Institut Gustave Roussy, 94805 Villejuif, France

Abstract *Chronic Myelomonocytic Leukemia (CMML) is a pathology of the elderly and the most frequent myelodysplastic - myeloproliferative neoplasm. The disease is associated with acquired somatic mutations. More than 30 genes have mutations in leukemic cells of 1 to 60% CMML patients. Analysis of 19 of these recurrently mutated genes of 264 patients identified at least one mutation in 95% of patients. We now extend this analysis to whole exome. We sequence leukemic cells of 29 patients in order to have a probability of 80% to detect a recurrent gene, mutated at a minimum of 10% frequency in CMML population. We declare that a gene is recurrently mutated if it is mutated in at least 2 patients out of 29. Tumor tissues were systematically compared to normal tissues from a same patient. We detected an average of 16 somatic mutations per patient in the sixteen first individuals. We found several genes already known to be frequently mutated in CMML and a novel gene mutated in 2 patients. We are currently exploring whether mutations in this later gene could have a functional impact in hematopoiesis.*

Keywords Chronic Myelomonocytic Leukemia, somatic mutations, exome sequencing, monocyte, T-lymphocyte, skin fibroblasts, *TET2*, *SRSF2*, *ASXL1*, alignment, variant calling, annotation.

1 Introduction

Chronic Myelomonocytic Leukemia is a clonal disease of hematopoietic stem cell affecting particularly the elderly. It is a rare pathology since the prevalence is $\simeq$ 1/100000 in France. The main diagnostic criteria is a persistent monocytosis. At the beginning of the disease, the symptoms are slight, mostly fewer, weight lost and fatigue. Patient death is due either to transformation in acute myeloid leukemia (30% of patients) or to consequences of cytopenias (haemorrhages and infections). The only curative therapy is allogenic hematopoietic stem cell transplant, which is rarely feasible because of patient age. Commonly used treatments include hydroxyurea and hypomethylating agents (azacytidine and decitabine). Hypomethylating drugs appear to delay disease evolution in around 40% of the severest forms of CMML [1]. CMML is characterized by chromosomic aberrations, genic expression anomalies and somatic mutations. In this study, we focused on somatic mutations. These mutations occur preferentially in particular genes, either all along the gene or at specific exons or bases. The most frequently mutated genes are epigenetic regulators, with *TET2* and *ASXL1* mutated in $\simeq$ 60% and $\simeq$ 40% of patients respectively, *IDH1*, *IDH2*, *EZH1*, *EZH2*, *UTX* and *DNMT3A*. Splicing factors are widely mutated with *SRSF2* altered in half of patients, *ZRSF2* and *U2AF1*. *CBL*, *NRAS*, *KRAS*, *JAK2*, *SH2B3*, *FLT3*, *KIT* and *MPL* are the most frequently mutated cytokinic signalisation regulators. Transcription factors are mutated as well, with *RUNX1* in 15% of patients and *NPM1*. Finally, *CUX1*, *SETPB1*, *BCOR*, *STAG2*, *RIT1*, *TP53* are recurrently mutated in CMML [2]. Analysis of 19 of these recurrently mutated genes in 264 patients identifies at least one mutation in 95% of patients (Itzykson, in press). To determine which other genes are recurrently mutated in CMML, we initiated exome sequencing of leukemic cells, using CD3+ lymphocytes or skin fibroblasts as genomic controls.

2 Materials and Methods

2.1 Cohort Establishment

First, we defined the number of patients N required for the study. Our aim is to detect recurrent genes mutated at a frequency f of at least 10%. Recurrent means here in at least two people. Among all the detected mutations, we further study the redondant ones. We chose 10% to focuse on pretty frequent mutations and to have a chance to either detect novel mutated genes or verify that there is no additional frequent mutated genes in CMML. The probability of observing 0 patient with mutation of gene G is (1-f)^{N}. The probability of observing exactly one patient with mutation of gene G is given by N.f.(1-f)$^{N-1}$. Thus, the probability p of observing twice or more the mutation of gene G is given by 1-[(1-f)^{N} + N.f.(1-f)$^{N-1}$]. We want p to be at least 80%. Thus, 1-[(1-f)^{N} + N.f.(1-f)$^{N-1}$] $\geq$ p. Therefore, the number N of CMML patients has been fixed at 29 in order to have a probability of 80% to get patients with ”recurrent” genes mutated at a frequency of at least 10% (Fig. 1). Note that the probability of detecting a mutation in a patient tends towards the probability that a patient present a mutation when the sample coverage is high and uniform.

We then defined the relevant material to use from patients. Most of our samples were taken at diagnostic time. All but one samples have been taken before the patient received any treatment for CMML. Tumor tissues were systematically compared to normal tissues from a same patient. Peripheral blood monocytes are used as leukemic cells and either T lymphocytes or skin fibroblasts as control cells. We tried to used as much as possible skin fibroblasts because there is no tumor content in these cells. However all patients did not agree to undergo biopsy. Thus, we have 9 skin fibroblasts and 20 T-lymphocytes as normal samples, which may indicate to us whether both cell types are good controls or not. Samples of patient's monocytes and skin fibroblasts were banked for whole-exome sequencing after patients provided written informed consent.

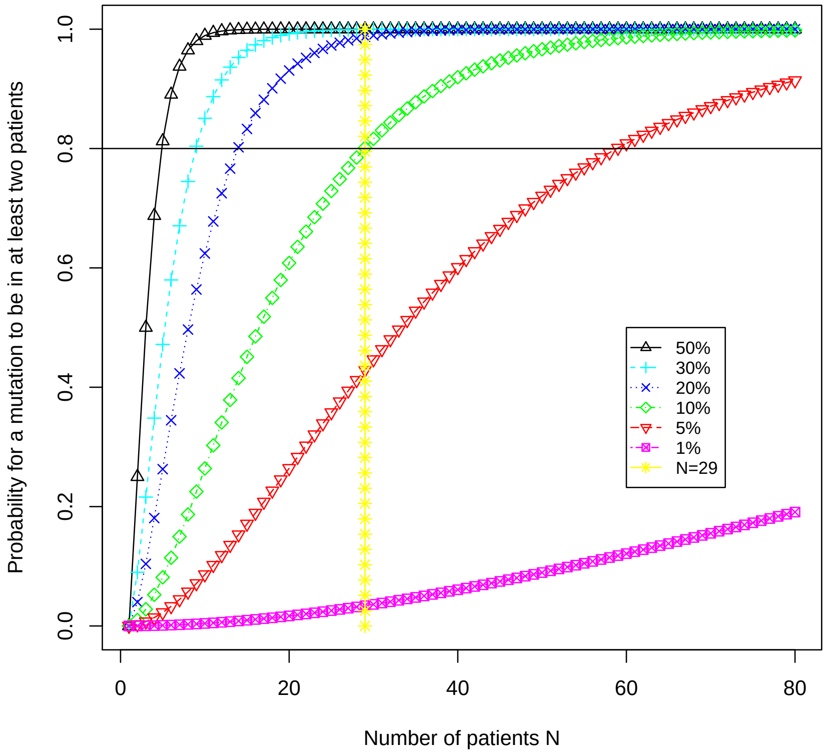


Figure 1. Probability of having at least two patients with a mutation of 50, 30, 20, 10, 5 or 1% frequency.

2.2 Exome Analysis

The exomes of tumor-normal samples were captured using the Agilent SureSelect Human All Exon kit v2 (46Mb), v3 (50Mb) or v4 (70Mb and 50Mb) and Illumina TruSeq v2 (62Mb) Target Enrichment kit according to the manufacturer's instructions. Captured DNA was run on the Illumina HiSeq 2000 platform with version 8 flow cells to generate 75 or 100 base paired-end reads. We require a minimum mean coverage of 50x for each sample. We have results for sixteen patients at the moment. The mean coverage of each base in the targeted regions is 95.69 (sd: 18.91). In average, the samples have a coverage $\geq$ 10x for 75.49% (sd: 8.69) of the bases.

Reads were aligned to the reference genome hg19 (NCBI human genome assembly build 37) using Burrows-Wheeler Aligner [4]. PCR duplicates were removed by Picard (http://picard.sourceforge.net). Local realignement of reads and base quality score recalibration were performed by Genome Analysis Toolkit [5]. VarScan [6] was used in the somatic mode to detect variants between normal and tumor sample of each patient. We filtered variants on the p-value of the Fisher Exact test performed in VarScan, which compares the proportion of variants between normal and tumor samples. This allows to eliminate most of germline variants and sequencing errors. Then, we annotated variants using Annovar [7]. We first removed variants present in dbSNP [8] version 129 since it is considered as the last "clean" database, *i.e.* it does not contain mutations. Then variants in non exonic regions and synonymous SNVs were deleted. Finally, we cured manually the variants on Integrative Genomics Viewer [9] (IGV).

3 Results

3.1 Overview of Somatic Mutations in CMML

The analysis of the sixteen first patients shows that 7225 variants remain after filtering on p-value, *i.e.* an average of 451,56 variants per patient (sd: 358,96). The deletion of variants present in dbSNP129 dropped this number to 2971, *i.e.* an average of 185,69 variants per patient (sd: 80,27). This number decreases to 1714, *i.e.* an average of 107,13 variants per patient (sd: 61.19) when removing loss of heterozygoty. We identified 381 variants in exonic and splicing regions. The two main categories are non synonymous SNVs and synonymous SNVs with $\simeq$ 50% and $\simeq$ 20% of variants respectively. The mutation types of these variants are reported Fig. 2. We removed synonymous SNV and performed a manual check on IGV. In total, we obtained 257 variants, *i.e.* an average of 16.06 (sd 5.99) somatic mutations per patient.

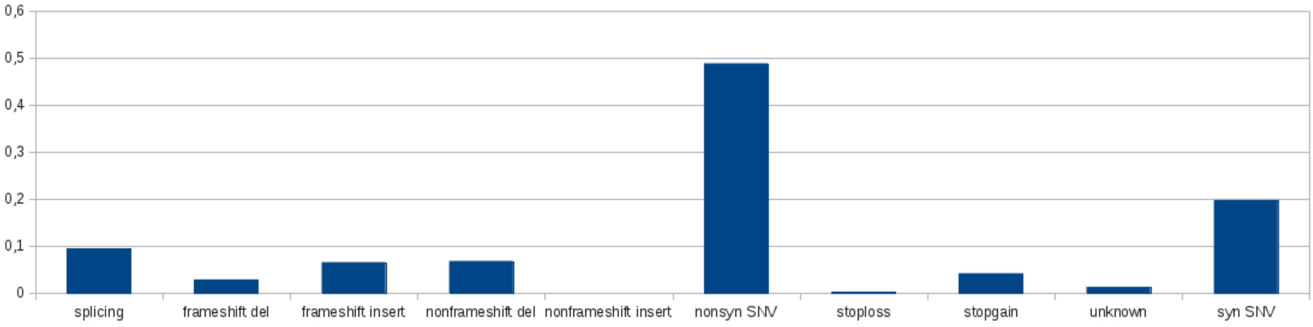


Figure 2. Mutation types in CMML.

Among the 257 detected variants, there are 203 SNPs (78.99%) and 54 INDELs. The SNPs are composed of 158 (77.83%) transitions (Ti) and 45 transversions (Tv). The ratio Ti/Tv $\simeq$ 3.51, which is expected to be [2.8;3] [3]. Somatic variants occur predominantly at G:C bases. Transitions G→A and C→T are the two most prevalent categories with 62 variants (30.54%) and 60 variants (29.56%) respectively (see Fig. 3 on the following page).

We found several genes implicated in CMML frequently mutated in our cohort (see Fig. 4 on the next page). We also detected two "novel" genes mutated in 2 patients. We further studied the most interesting candidate that we validated *via* Sanger sequencing: the two mutations identified were present in patient monocytes but

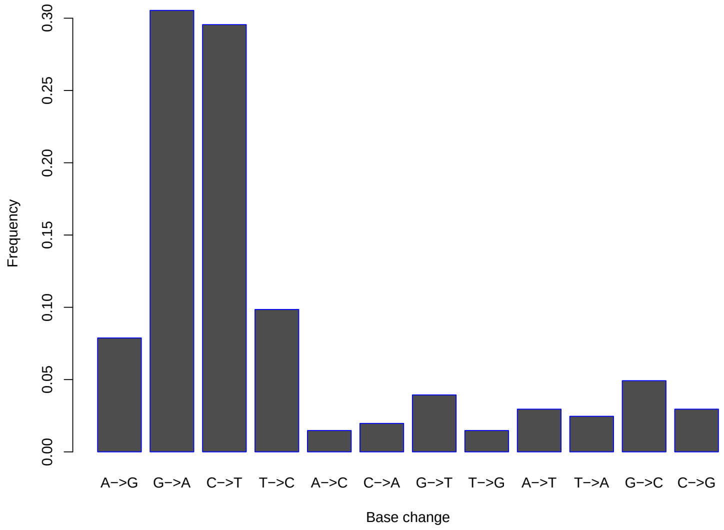


Figure 3. Frequency of base changes in 16 CMML patients.

absent in skin fibroblasts. This gene is mutated in a series of other tumor types, although its precise role (driver or passenger) is unknown. Our first PCR assay attempting to detect the gene expression in hematopoietic stem cells and mature cells was negative, suggesting that recurrent mutations in this large gene may have limited pathophysiological impact in hematopoietic cells.

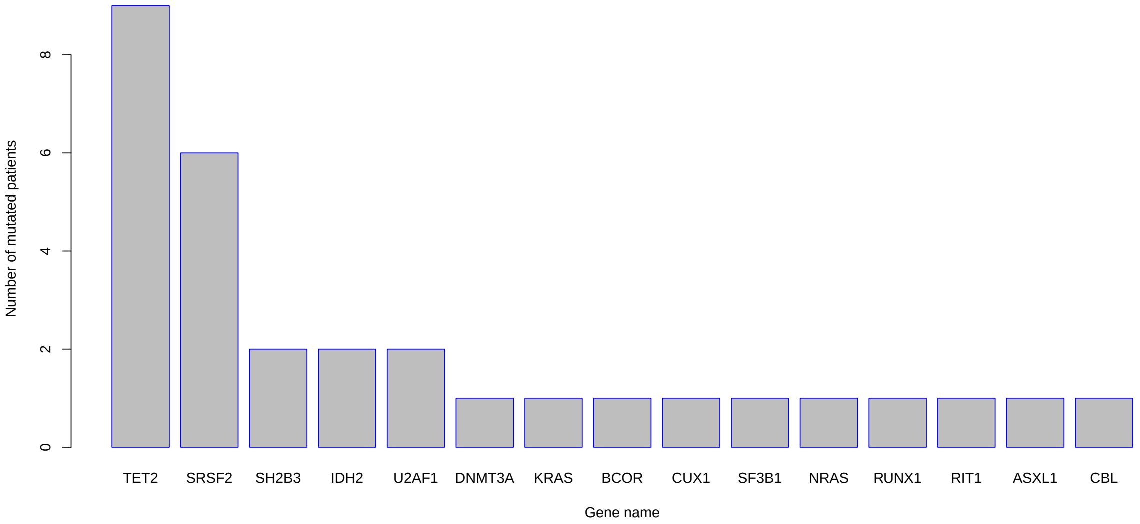


Figure 4. Known genes of CMML mutated in our cohort.

3.2 Somatic Mutations Validation

Validation of all putative mutations is currently performed using AMPLI SEQ technology and deep sequencing life technologies Ion Torrent PGM (multiplex PCR on chip 318). We plan to achieve a coverage of 1000x. Novel recurrently mutated genes will be in addition manually confirmed by Sanger sequencing.

4 Conclusion

Exome sequencing of both monocytes and T-lymphocytes or skin fibroblasts allowed us to detect acquired mutations. On average, we detected 16 somatic mutations per patient. We found not only genes frequently mutated in CMML like *TET2*, *SRSF2*, *IDH2* but also novel genes recurrently mutated in CMML. We study currently the functional impact of our most interesting candidate. We hope to get more candidates once our cohort is fully sequenced. Furthermore, we want to study the impact of time and treatment over mutational clones. For that, we sequence exome of leukemic cells at different times before and/or after treatment for a same patient.

Acknowledgements

This work was supported by the Ligue Nationale Contre le Cancer (Label), Agence Nationale de la Recherche, Institut National du Cancer. The Unite Mixte de Recherche 1009 equipment was supported by the Association pour la Recherche sur le Cancer.

References

[1] T. Braun, R. Itzykson, A. Renneville, B. de Renzis, F. Dreyfus, K. Laribi, K. Bouabdallah, N. Vey, A. Toma, C. Recher and others, Molecular predictors of response to decitabine in advanced chronic myelomonocytic leukemia: a phase 2 trial. *Blood*, 118:3824-3831, 2011.

[2] R. Itzykson, N. Droin, and E. Solary, Les progrès récents dans la leucémie myélomonocytaire chronique. *Hématologie*, 18:24-36, 2012.

[3] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. Del Angel, M.A. Rivas, M. Hanna, and others, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43:491-498, 2011.

[4] H. Li and R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25:1754-1760, 2009.

[5] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and others, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data *Genome research*, 20:1297-1303, 2010.

[6] D.C. Koboldt, Q. Zhang, D.E. Larson, D. Shen, M.D. McLellan, L. Lin, C.A. Miller, E.R. Mardis, L. Ding, and R.K. Wilson, VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing *Genome Research*, 22:568-576, 2012.

[7] K. Wang, M. Li, and H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data textitNucleic acids research, 38:e164-e164, 2010.

[8] ST Sherry, M-H Ward, M. Kholodov, J. Baker, L. Phan, EM Smigielski and K. Sirotkin, dbSNP: the NCBI database of genetic variation *Nucleic acids research*, 29:308-311,2001.

[9] J. T Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S Lander, G. Getz, and J. P Mesirov, Integrative genomics viewer *Nature biotechnology*, 29:24-26, 2011.

Insyght: using symbols to visualize homologies, conserved syntenies and genomic insertions across multiple genomes

Thomas Lacroix[1], Valentin Loux[1], Annie Gendrault[1], Mark Hoebeke[2] and Jean-François Gibrat[1]

[1] Mathématique, Informatique et Génome, INRA, Jouy-en-Josas, 78352 France
{thomas.lacroix, valentin.loux, annie.gendrault, jean-francois.gibrat}@jouy.inra.fr

[2] CNRS, UPMC, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France
mark.hoebeke@sb-roscoff.fr

Abstract *Insyght proposes a new way to explore the landscape of conserved and idiosyncratic genomic regions across multiple genomes and their rearrangements throughout evolution. Its unique display consists of a symbolic representation tightly integrated with a proportional view. The symbols highlight a region of interest and provide legibility while the proportional view simultaneously allows grasping genomic locations and complex rearrangements scattered across the genomes and occurring at different scales. A second type of display is dedicated to the analysis of the presence, absence, or multiple copies of a given set of homologs. A functionality based on filters has been implemented to facilitate the retrieval of genes of interest and allow the formulation of relevant biological questions, such as finding niche-specific or core genome genes that match a few particular functions or biological processes. Our public dataset currently consists of 389 prokaryotes genomes. Alternatively, a virtual machine can be downloaded and installed locally to visualize private data. It contains a pre-installed version of the pipeline, database and visualisation tool. Insyght is suitable for a variety of analyses: genome-wide inference of gene function, detection of evolutionary events, phylogenetic profiling and investigation of the core genome or niche-specific genes. It is freely available at http://genome.jouy.inra.fr/Insyght/*

Keywords homology browser, conserved synteny, multi-genome visualisation

1 Introduction

Les génomes subissent des réarrangements de différentes natures lors des processus évolutifs: translocation, duplication, fusion, etc... D'un point de vue de la génomique comparée, chaque génome peut être considéré comme une succession de régions conservées intercalées par des régions idiosyncratiques. Les technologies de séquençage haut débit fournissent une grande quantité de données aux biologistes qui ont besoins d'outils pour les aider à annoter les gènes de manière efficace et rapide à l'échelle du génome. La conservation de l'ordre des gènes peut permettre d'assigner les fonctions d'un ensemble de gènes simultanément ou fournir des indices concernant la fonction des protéines hypothétiques [1,2]. De nombreux outils d'exploration de synténies existent, et les paradigmes de visualisation dans ce domaine sont variés: le dot plot [3,4,5,6], l'idéogramme [7,8], la vue en trapézoïde [9,10,11], ou la représentation symbolique [12,13,14]. Parmi les défis posés par l'analyse des synténies et homologues, on peut noter la visualisation des réarrangements qui sont répartis sur les génomes et se produisent à différentes échelles, ou la navigation parmi une quantité de donnée abondante et multi-dimensionnelle (coordonnée génomique, plusieurs génomes comparés, plusieurs homologues par comparaison,...).

2 Design et fonctionnalités principales

Insyght est un outil de visualisation qui permet d'analyser les homologies, les synténies conservées et les régions génomiques idiosyncratiques à l'échelle de plusieurs organismes. Sa caractéristique principale est l'association d'une représentation symbolique (Figure 1-A) à une représentation proportionnelle (Figure 1-B). Cette combinaison originale de paradigme visuel facilite la navigation exhaustive de données d'homologies complexes. L'utilisateur peut interagir avec divers symboles qui représentent les évènements évolutifs:

homologues, synténies, régions génomiques insérées. Ces symboles sont étroitement intégrés avec une vue proportionnelle où les mêmes évènements sont représentés selon leurs coordonnées génomiques. Dans la vue proportionnelle, les régions génomiques homologues sont jointes par des trapézoïdes. La représentation symbolique améliore la lisibilité parmi le grand nombre d'événements évolutifs et permet de naviguer parmi les multiples copies d'homologues. La représentation proportionnelle permet de localiser les réarrangements complexes dispersés dans le génome et se produisant à différentes échelles. L'utilisateur peut interagir avec les symboles à l'aide d'un menu contextuel pour, par exemple, afficher ou masquer les gènes d'une synténie ou trouver des gènes d'intérêt. Le zoom et la navigation peuvent être synchronisés entre les résultats ce qui permet d'analyser plusieurs génomes en parallèle.

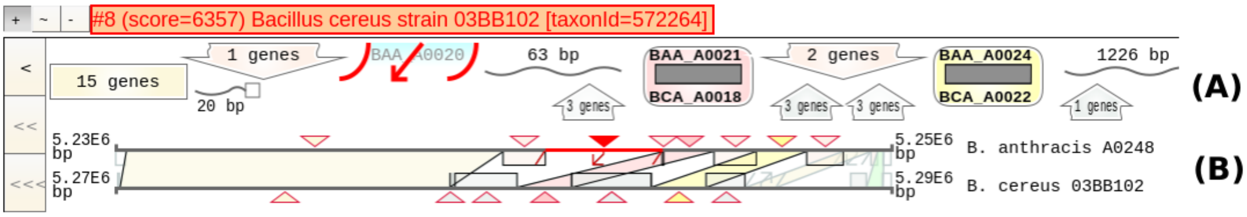
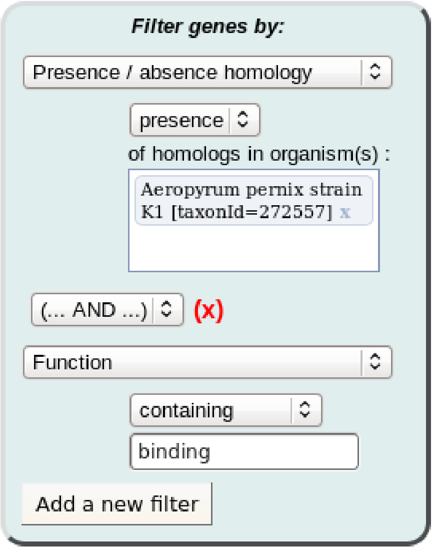


Figure 1. Représentation symbolique (A) et proportionnelle (B) de la vue d'organisation génomique.

Une deuxième vue est consacrée à l'analyse exhaustive des homologies d'un jeu de gènes d'intérêt. Une fonctionnalité de recherche par combinaison de filtres (Figure 2) a été implémentée pour faciliter la constitution de groupes de gènes significatifs d'un point de vue biologique. Les opérateurs booléens logiques d'intersection (AND) ou d'union (OR) permettent de combiner différents types de filtres: présence / absence d'homologues, coordonnées génomiques, identifiants, fonctions, processus biologiques, produits, localisation cellulaire, ou numéro EC. Par exemple, il est possible de formuler des requêtes combinées qui permettent de trouver les gènes spécifiques ou partagés au sein d'une espèce correspondant à un processus biologique particulier. La vue d'analyse des homologues ressemble à un tableau où les colonnes sont les gènes sélectionnés et les lignes sont les espèces comparées (Figure 3). Ainsi l'utilisateur peut visualiser la présence, l'absence, ou les multiples copies d'homologues et détecter par exemple les espèces avec des pertes de fonction ou des familles de gènes abondantes. Les gènes sont colorés en fonction de la synténie à laquelle ils appartiennent. D'autre fonctionnalités ont été implémentées, tel que trier le tableau de résultat selon divers critères et échelles ou visualiser l'emplacement des gènes sur les génomes. Les deux vues, organisation génomique et tableau des homologues, sont interconnectées et il est possible de passer de l'une à l'autre.



Filter genes by:

Presence / absence homology

presence

of homologs in organism(s) :

Aeropyrum pernix strain K1 [taxonId=272557] x

(... AND ...) (x)

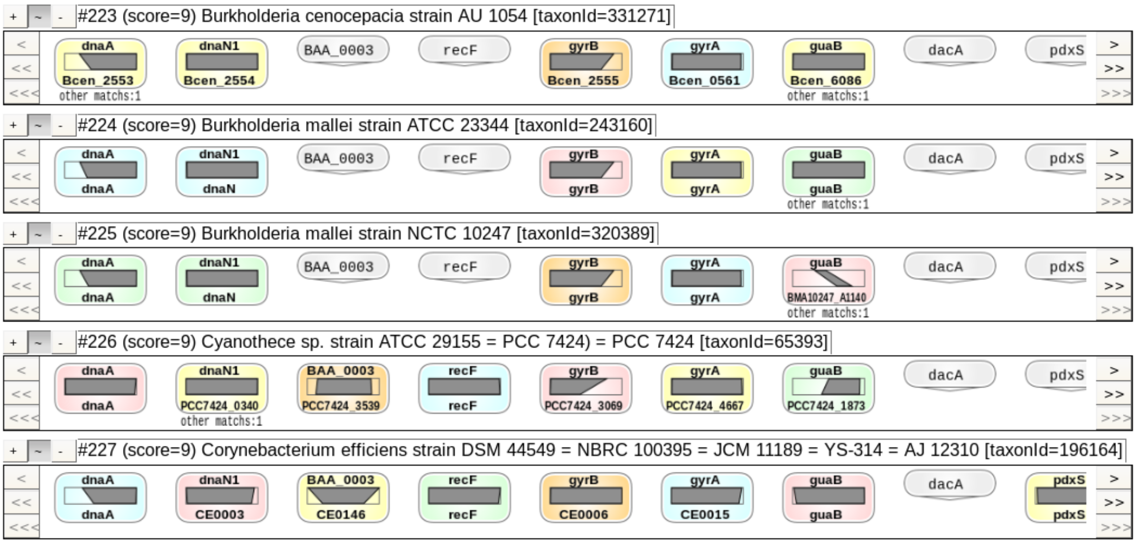
Function

containing

binding

Add a new filter

Figure 2. Fonctionnalité de recherche de gènes (combinaison de filtres).



#223 (score=9) Burkholderia cenocepacia strain AU 1054 [taxonId=331271]

#224 (score=9) Burkholderia mallei strain ATCC 23344 [taxonId=243160]

#225 (score=9) Burkholderia mallei strain NCTC 10247 [taxonId=320389]

#226 (score=9) Cyanothece sp. strain ATCC 29155 = PCC 7424) = PCC 7424 [taxonId=65393]

#227 (score=9) Corynebacterium efficiens strain DSM 44549 = NBRC 100395 = JCM 11189 = YS-314 = AJ 12310 [taxonId=196164]

Figure 3. Vue tableau des homologues

389 génomes complets procaryotes ont été intégrés dans la base de données PostgreSQL à ce jour. Le pipeline s'appuie sur Genome Reviews, BLAST, le bi-directional best hit pour inférer l'orthologie, et la programmation dynamique pour déterminer les synténies. Toutes les données et méthodes utilisées par le pipeline sont publiques ou ont été publiées par leurs auteurs. L'interface est une application web développée en GWT et HTML5. Pour analyser des données privées, une machine virtuelle peut être téléchargée qui contient une version pré-installée du pipeline, de la base de données et de l'application web de visualisation.

3 Conclusions

Insyght (http://genome.jouy.inra.fr/Insyght) propose une représentation visuelle et une navigation originale des synténies et homologies qui ouvrent une perspective nouvelle en ce qui concerne diverses analyses biologiques classiques: annotation de la fonction des gènes à l'échelle du génome, détection des évènements d'évolutions (par exemple transfert horizontal), profilage phylogénétique, et analyse de gènes niche-spécifiques ou core-génome. L'analyse des gènes dans le contexte d'espèces proches ou distantes phylogénétiquement est un besoin identifié par les biologistes [15,16]. Insyght permet de constituer un jeu de gènes qui satisfait plusieurs critères hétérogènes et d'analyser les gènes candidats. La vue du tableau d'homologues offre une approche exhaustive et simple pour étudier les homologies abondantes. La vue génomique permet l'identification d'évènements évolutifs et améliore la lisibilité parmi le grand nombre de régions synténiques et idiosyncratiques.

Remerciements et financements

Les auteurs souhaitent remercier Dr Philippe Bessières et Dr Catherine Juste pour leur commentaires constructifs. Ce travail a été financé par l'Agence Nationale de la Recherche [ANR-PDR-080124-03-01].

References

[1] M. Huynen, B. Snel, W. Lathe, and P. Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res*, 10, 1204-1210, 2000.

[2] X.H. Zheng, F. Lu, Z.Y. Wang, F. Zhong, J. Hoover, and R. Mural, Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. *Bioinformatics*, 21, 703-710, 2005.

[3] T. Davidsen, E. Beck, A. Ganapathy, R. Montgomery, N. Zafar, Q. Yang, R. Madupu, P. Goetz, K. Galinsky, O. White, *et al.,* The comprehensive microbial resource. *Nucleic Acids Res*, 38, D340-345, 2010.

[4] J. Blom, S.P. Albaum, D. Doppmeier, A. Puhler, F.J. Vorholter, M. Zakrzewski, and A. Goesmann, EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics*, 10, 154, 2009.

[5] E. Courcelle, Y. Beausse, S. Letort, O. Stahl, R. Fremez, C. Ngom-Bru, J. Gouzy, and T. Faraut, Narcisse: a mirror view of conserved syntenies. *Nucleic Acids Res*, 36, D485-490, 2008.

[6] C. Soderlund, M. Bomhoff, W.M. and Nelson, SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res*, 39, e68, 2011.

[7] A.U. Sinha, J. and Meller, Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, 8, 82, 2007.

[8] D.R. Riley, S.V. Angiuoli, J. Crabtree, J.C. Dunning Hotopp, and H. Tettelin, Using Sybil for interactive comparative genomics of microbes on the web. *Bioinformatics*, 28, 160-166, 2012.

[9] S.J. McKay, I.A. Vergara, J.E. and Stajich, Using the Generic Synteny Browser (GBrowse_syn). *Curr Protoc Bioinformatics*, Chapter 9, Unit 9 12, 2010.

[10] D. Vallenet, L. Labarre, Z. Rouy, V. Barbe, S. Bocs, S., Cruveiller, A. Lajus, G. Pascal, C. Scarpelli, and C. Medigue, MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res*, 34, 53-65, 2006.

[11] Y.Wang, H.Tang, J.D. Debarry, X. Tan, J. Li, X. Wang, T.H. Lee, H. Jin, B. Marler, H. Guo, *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*, 40, e49, 2012.

[12] T. Derrien, C. Andre, F. Galibert, C. and Hitte, AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. *Bioinformatics*, 23, 498-499, 2007.

[13] M. Muffato, A. Louis, C.E. Poisnel, and H. Roest Crollius, Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, 26, 1119-1121, 2010.

[14] F. Lemoine, B. Labedan, and O. Lespinet, SynteBase/SynteView: a tool to visualize gene order conservation in prokaryotic genomes. *BMC Bioinformatics*, 9, 536, 2008.

[15] K.E. Nelson, D.E. Fouts, E.F. Mongodin, J. Ravel, R.T. DeBoy, J.F. Kolonay, D.A. Rasko, S.V. Angiuoli, S.R. Gill, I.T. Paulsen, *et al.* Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen Listeria monocytogenes reveal new insights into the core genome components of this species. *Nucleic Acids Res*, 32, 2386-2395, 2004.

[16] T.D. Read, G.S. Myers, R.C. Brunham, W.C. Nelson, I.T. Paulsen, J. Heidelberg, E. Holtzapple, H. Khouri, N.B. Federova, H.A. Carty, *et al.* Genome sequence of Chlamydophila caviae (Chlamydia psittaci GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. *Nucleic Acids Res*, 31, 2134-2147, 2003.

Functional divergence in protein families: a co-variation analysis

Julien PELE, Matthieu MOREAU and Marie CHABBERT
LABORATOIRE BNMI, UMR CNRS 6214 – INSERM 1083, Faculté de Médecine, 3 rue Haute de reculée,
49045, Angers, France
marie.chabbert@univ-angers.fr

Abstract *We compare different co-variation methods for their ability to mine co-varying positions related to functional divergence in a protein family. The methods are tested on several sets of G-proteins-coupled receptors (GPCRs). Co-variation methods are based on a substitution matrix, the χ^2 score, mutual information or a perturbation algorithm. For each method, the detailed analysis of the top pairs is performed. A graphical tool that we have developed helps the interpretation of the data. We show that, to mine groups of co-varying positions related to functional divergence, the co-variation methods must privilege positions with intermediate entropy. The OMES (Observed Minus Expected Squared) method, proposed by Fodor and Aldrich (2004), is especially adapted for this purpose.*

Keywords Sequence analysis, sequence co-variation, protein family, protein evolution, GPCR.

1 Introduction

Positions in the multiple sequence alignment (MSA) of a protein family are not independent and their co-variation may reflect structural and/or functional constraints. Various methods have been developed to analyze these co-variations (reviewed in [1]) and may provide different information. Here, we compare six co-variation methods on sets of G-protein-coupled receptors (GPCRs) [2, 3] to get insights into specificity-determining positions. We have developed a graphical tool that helps the interpretation of the data.

2 Methods

The sequence conservation at each position is measured by sequence entropy. The co-variation methods [1, 4] tested are: (1) the McLachlan Based Substitution Correlation (McBASC) method, (2) the Observed Minus Expected Squared (OMES) method adapted from the χ^2 test, (3-5) the mutual information (MI) method and the MIp and MINT corrections, based on the probability of joint occurrence of events and (6) the Explicit Likelihood of Subset Covariation (ELSC), based on a perturbation algorithm. We focus on the detailed analysis of a limited number of top pairs. To interpret the results, the top 25 pairs are sorted in descending order and appended to form a MSA. The sequences in the alignment are sorted according to sub-families. At each position, the amino acids are visualized with a color code, to relate co-variation scores to the amino acid distribution at the corresponding positions. The most frequent and the second most frequent amino acids are blue and yellow, respectively. Other amino acids are white.

3 Results

The MSA of the sorted top pairs reveals three coloring schemes. In the mono scheme, the blue color of the most frequent pair dominates. In the mixed one, the white color, mixed with blue and yellow, dominates. The third pattern corresponds to a dual (blue/yellow) coloring scheme. The mono scheme is observed with McBASC. The mixed scheme is observed for MI and, to a lesser extent, MINT. The most typical example of dual scheme is observed with OMES, but it is also the case for ELSC and MIp. These schemes can be related to the entropy of the top pairs whose average value increases in the following order:

$$\text{McBASC} < \text{OMES} < \text{ELSC} \approx \text{MIp} < \text{MINT} < \text{MI}.$$

Fig. 1 shows typical results, obtained from the analysis of a GPCR subset that includes about 100 receptors from the somatostatin/opioid (SO), the chemotactic (CHEM) and the purinergic (PUR) sub-families that are evolutionary related by divergence [2]. OMES reveals a split between the SO/CHEM and the PUR sub-families, for 22 out of the 25 pairs. A similar pattern is obtained with the ELSC and the MIp methods for 13 and 10 pairs, respectively. The detailed analysis of the top pairs indicates that OMES privileges pairs involving a residue with a high connectivity [5]. In particular, in the example shown, 11 out of the top 25 pairs involve the As/Asp mutation specific of the divergence of PUR receptors.

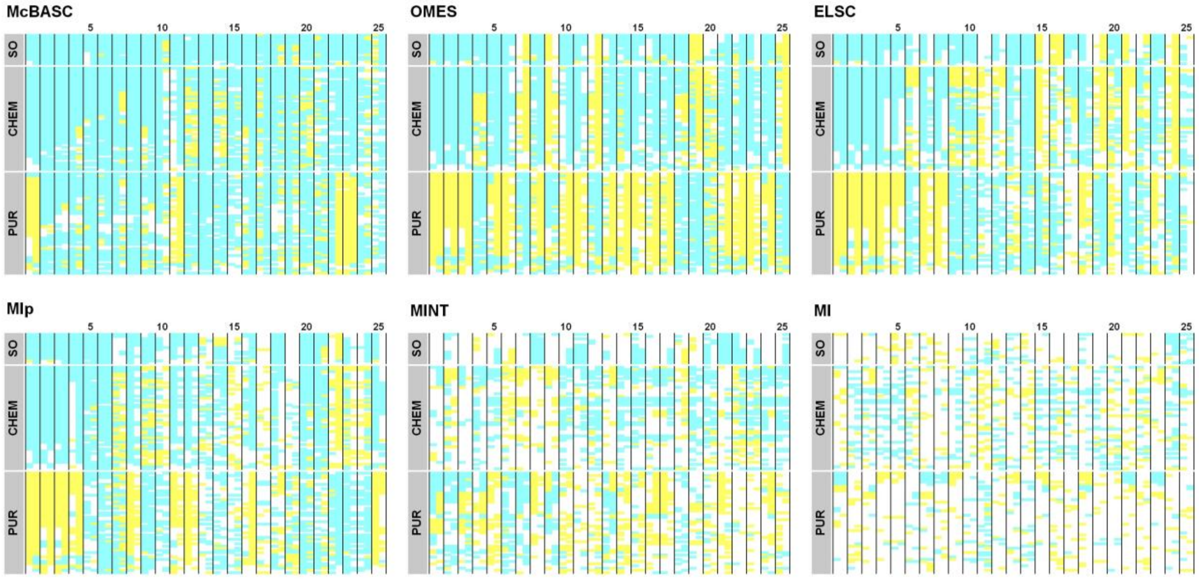


Figure 1. Multiple alignments of the top 25 pairs of co-varying positions in the MSA of the SO, CHEM and PUR sub-families. The most frequent and second most frequent residues are colored blue and yellow, respectively.

4 Conclusions

The relationship between co-variation scores and residue conservation depends on the algorithms that may favor conserved or variable positions [4]. We searched for a method that privileges pairs of co-varying positions in link with sub-family divergence. This is the case for the OMES, ELSC and MIp methods that select pairs with an intermediate conservation level. In particular, the OMES method appears to be very adapted to analyze co-variations related to functional divergence within a protein family.

Acknowledgements

This work was supported by a grant from ANR (ANR-11-BSV2-026) and by institutional grants from CNRS, INSERM and University of Angers. We thank NEC Computers Services SARL (Angers, France) for the kind provision of a multiprocessor server. JP and MM were supported by fellowships from Conseil Général de Maine-et-Loire and from CNRS, respectively.

References

[1] D. de Juan, F. Pazos and A. Valencia, Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14:249-261, 2013.

[2] J. Deville, J. Rey and M. Chabbert, An indel in transmembrane helix 2 helps to trace the molecular evolution of class A G-protein-coupled receptors. *Journal of Molecular Evolution*, 68:475-489, 2009.

[3] J. Pele, H. Abdi, M. Moreau, D. Thybert and M. Chabbert, Multidimensional scaling reveals the main evolutionary pathways of class A G-protein-coupled receptors. *PloS one*, 6:e19094, 2011.

[4] A.A. Fodor and RW Aldrich, Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, 56:211-221, 2004.

[5] R. Merkl and M. Zwick, H2r: identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments. *BMC Bioinformatics*, 9:151, 2008.

Comparison of Numerical Resolution Methods for Biological Kinetic Models

Merdan SARMIS[1, 2], Jean-Marie BOUTEILLER[1], Serge BISCHOFF[1], Olivier HAEBERLE[2], and Michel BAUDRY[1]

[1] RHENOVIA PHARMA SAS, 20c rue de Chemnitz, Technopole – Mer Rouge Plaza, 68200, Mulhouse, Cedex, France
merdan.sarmis@rhenovia.com

[2] LABORATOIRE MIPS, EA 2332, 12 rue des Frères Lumière, 68093, Mulhouse, Cedex, France
olivier.haeberle@uha.fr

Abstract *Models of biological synaptic receptors are generally based on systems of kinetic reactions. These reactions are translated into ordinary differential equations (ODEs), which are computationally resolved by a numerical ODE solving algorithm. Computation performance, defined as the rapid convergence of the algorithm to a numerical solution of the ODE system, critically depends on the choice of the appropriate ODE algorithm. In this paper we compare several ODE algorithms using two types of kinetic models, and determine their performances in order to provide a benchmark for these models. The benchmark will facilitate the choice of the most efficient algorithm for a given kinetic model with a minimum number of tests. Our results provide a tool for identifying optimal potential ODE solvers for any type of models under various experimental conditions. This comparison also underscored the complexity of biological kinetic models and how it could interfere with ODE solver performance. Despite these challenges, we were able to select the best solvers for any synaptic receptor kinetic models described with a bilinear system without any a priori information on the ODE solver structure. Based on our results, we recommend using the solver named LSODE with a common tolerance set for bilinear synaptic receptors kinetic models.*

Keywords Benchmark, ODE, Numerical Resolution Method, Kinetic Model, Synaptic Receptor, Bilinear System.

1 Introduction

The main goal of systems biology consists in providing a quantitative and integrative description of living organisms by using simulation of complex and interconnected models of metabolic networks and signaling pathways. These complex processes are described by systems of biochemical reactions and quantitatively analyzed by corresponding sets of mathematical equations. In order to analyze the temporal evolution of the systems of reactions, series of ordinary differential equations (ODEs) that relate reactant concentrations and elementary rate constants to changes in product concentrations as a function of time need to be resolved. ODEs are computationally solved with ODE solver algorithms, providing numerical solutions for the temporal changes of the various variables of the biological systems. Over the years, mathematicians have developed numerous ODE algorithms. However, each system or model is another problem to resolve and each algorithm differ with its rapidity accuracy to solve an ODE system. This is due in part to the fact that, given the wide feature range of ODE systems, such as stiff, non-stiff, linear, non-linear, etc., different ODE solvers produce relatively similar results, but with a wide range of performance, which is generally defined as the combination of the speed to reach a stable numerical solution and the accuracy of the final solution. It is therefore difficult to determine which algorithm to use to solve particular sets of ODE. In fact, as it described by Rice [1] and presented on Figure 1, algorithm selection is a complex problem, which depends on many criteria.

According to Rice (Fig. 1), in order to select the most appropriate algorithm, the user needs to know the model (problem space), the tool (feature space) and the ODE solver structure (algorithm space) in addition to user preferences (criteria space). Since users can be biologists, modelers, tool developers or mathematicians, they will have their own expertise, and only a small number of scientists will have expertise in all these domains. How then could a user make the right decision without expertise in all such domains? In our efforts to develop a simulation platform for hippocampal glutamatergic synapses [2] and its integration into complex neuronal networks, we repeatedly face this problem. As this project was initiated several years ago using the Java programming language, we propose to compare 8 Java-based ODE solver algorithms and to determine the most appropriate one for bilinear synaptic receptor kinetic models, depending on the properties of the system under consideration.

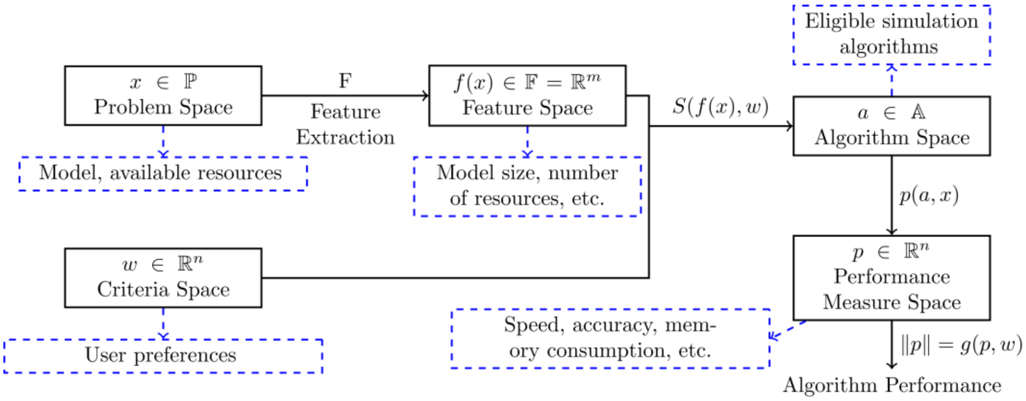


Figure 1. The algorithm selection problem as defined by Rice in 1976. As presented in this scheme, users need to know the model, tools and algorithm to be able to select the right solver for a simulation.

The 8 solvers that we compare were categorized into 3 groups: implicit, explicit and hybrid (hybrid solvers combine implicit and explicit methods for each integration step) solvers. Our synaptic receptor kinetic models are usually bilinear systems.

All 8 evaluate solvers have variable step-size, except for the selected reference solver, a fourth-order explicit Runge-Kutta ODE solver algorithm [3] (RK4). Explicit Runge-Kutta methods are stable algorithms, as long as the step-size remains small enough to avoid instabilities. Therefore, we selecte the RK4 algorithm solver with a constant step-size of 0.5 µsec as our reference to ensure that the solution of the simulation would remain stable and accurate. The second ODE solver is the Runge-Kutta Fehlberg [4] (RKF), a fourth/fifth-order explicit RK scheme. TR-BDF2 [5], an implicit solver, is composed of a Trapezoidal rule and a second-order Backward Differential Formula (BDF scheme). The Rosenbrock solver [6] and IMEX (for Implicit-Explicit) [7] are both hybrid fourth-order Runge-Kutta. CVODE [8], a widely used ODE solver, integrates two schemes: a variable order (1 to 12) explicit Adams-Moulton for non-stiff systems, and a variable order (1 to 5) implicit BDF for stiff systems. These two schemes are respectively called CVODE ADAMS and CVODE BDF. The last ODE solver is LSODE [9], which has the same Adams-Moulton and BDF schemes as CVODE. This last solver adds stiffness detection, as written by Linda R. Petzold [10]. With this stiffness detection, LSODE starts with the Adams-Moulton scheme and switches to the BDF scheme if the system becomes stiff, and vice-versa.

2 Material and Methods

For benchmarking, we use Rhenovia's biosimulation platform, RHENOMS™, and two models of synaptic receptors/channels. The first model we use in this benchmark is a model of the NMDA (NR1/NR2A) receptors [11], which we qualify as slow, since the kinetics of receptor activation are relatively slow (50-250 msec). The second model is a N-type voltage-dependent calcium channel (VDCC), which is a relatively fast model (0-5 msec for activation) and is based on the Hodgkin-Huxley formalism [12]. As the goal of this paper was not to discuss in details the models we used, we invite interested readers to read the appropriate references cited in this article.

We activate each model with two different protocols: 1) P1 (for Protocol 1) is a single event lasting 1 msec, and 2) P2 (for Protocol 2) is the same event repeated 4 times with 10 msec of interval. The duration of each simulation is set at 500 msec. The two protocols are tested with two sets of tolerance: 1) a common set of tolerance with relative tolerance fixed at $1e^{-3}$ (Rtol) and absolute tolerance fixed at $1e^{-6}$ (Atol), 2) and a set of *restrictive* tolerance with Rtol=$1e^{-6}$ and Atol=$1e^{-9}$. These two sets of tolerances are called *Tol1* and *Tol2*, respectively.

To assess performance, we use 4 criteria: execution time, number of points (memory consumption), Mean Square Error (MSE) and Normalized Root Mean Square Deviation (NRMSD) between the results produced by a solver and those generated by the reference solver (RK4). These 4 performance criteria are used to quantify the algorithms performance with a ||*p*|| norm, as illustrated in the algorithm selection problem diagram (Figure 1). The norm value provides a rapid comparison of the overall performance of the various algorithms, as it generates a single value for the performance of each algorithm. Algorithms yielding a small norm value will thus be considered as more efficient than others yielding a large value. Simulations are performed on a WorkStation with a LINUX (Ubuntu 10.04) operating system and an Intel® Xeon® CPU at 2.67 GHz frequency equipped with 12 Gbytes of RAM and the version 1.6 of Java installed.

3 Results and Discussion

In computational neuroscience, most models are stimulus-dependent and therefore bilinear. The major goal of our study is to benchmark the solvers implemented into RHENOMS and to provide a recommendation for their possible use for biological synaptic receptor kinetic models by users. Therefore, we will not discuss the details of the solvers or of the models.

3.1. NMDA Receptor Model

The NMDA type of glutamate receptor is a relatively slow (50-250 msec) ligand-gated channel. The model we use in this study was previously calibrated and validated [13] to fit a variety of experimental data. The ||*p*|| norm performances are depicted for P1 and P2 respectively on Fig. 2A and Fig. 2B. In order to find the best algorithm for this model we identify the solver(s) with the smallest norm values for the two stimulation protocols and the two tolerance sets.

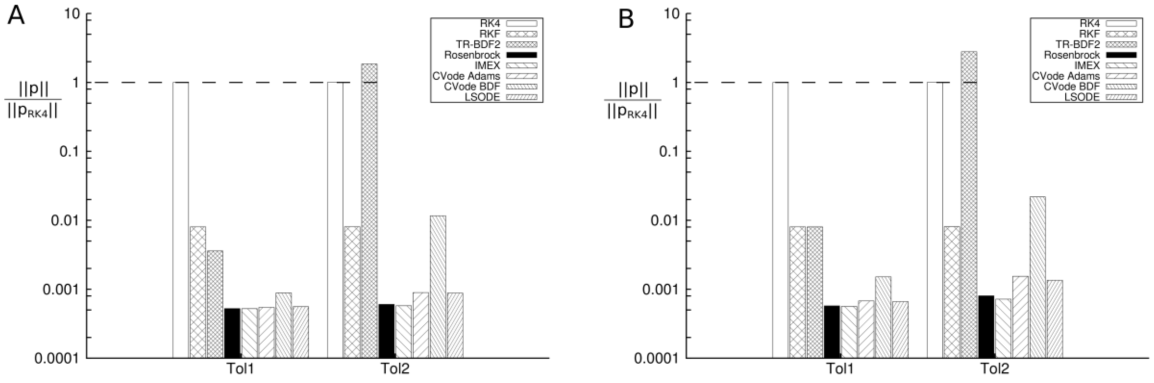


Figure 2. Normalized norm values for the 8 solver algorithms for the NMDA receptor model. A: Stimulate with P1. B: Stimulate with P2. All ||*p*|| norms are normalized with the reference RK4 $||p_{\mathrm{RK4}}||$ norm.

For P1 protocol, the algorithm performances differ not much between *Tol1* and *Tol2* tolerances except for the BDF scheme (TR-BDF2 and CVODE BDF solvers), where the differences are more than 10 fold on ||*p*|| performance. For P2 protocol, the performance differences are more pronounced with the last three algorithms between *Tol1* and *Tol2* parameters. As for P1 protocol, BDF scheme performances deteriorate significantly with *Tol2*. For both stimulation protocols, the best solver is IMEX for both tolerance parameters, while the less performing algorithm is RKF for *Tol1* and TR-BDF2 for *Tol2* for both protocols.

As a hybrid solver provides the best performance for the NMDA receptor model with all simulations, we cannot make general conclusions regarding model stiffness or input dependency of the model. Nevertheless, we are able to identify which solver is optimal (hybrid IMEX solver) to use with this model.

3.2. N-type VDCC model

In general, it has proven difficult to model voltage-dependent calcium channels (VDCC), due to the existence of a tail current at the end of the plateau current. For our comparative studies, we choose to simulate the N-Type VDCC, a high voltage-activated calcium channel. This channel is very fast, 0-5 msec for activation. The VDCC model we use is parameterized and validated to fit Jaffe and Poirazi results [14,15]. The $||p||$ norm performances are depicted for P1 and P2 respectively on Fig. 3A and Fig. 3B.

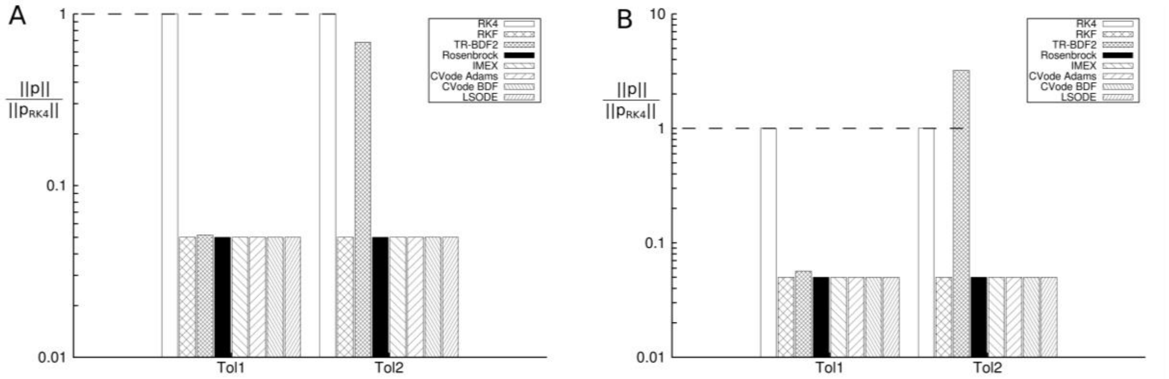


Figure 3. Normalized norm values for the 8 solver algorithms for the N-type VDCC model. A: Stimulate with P1. B: Stimulate with P2. All $||p||$ norms are normalized with the reference RK4 $||p_{\text{RK4}}||$ norm.

No significant differences are observed between all algorithms, except for TR-BDF2, which generated the worst performance value for both protocols and tolerance sets. Comparing the overall performances of the solvers, LSODE is the algorithm that provides the best performance for both stimulation protocols and tolerance sets. However, RKF performances are very close to that of JLSODE ($1e^{-6}$ difference). With the N-type VDCC model, we observe clearly that all solvers provide approximately the same performances, except for TR-BDF2 solver. Switching from *Tol1* to *Tol2*, TR-BDF2 solver performances exceeds RK4 values. Infact, a small relative tolerance value ($Rtol<1e^{-6}$) is not required to obtain accurate enough results with TR-BDF2 solver.

As all ODE solvers provide similar performances with the N-type VDCC model, we cannot conclude on model stiffness or input dependency for the model. Nevertheless, we are able to identify which solver is optimal to use with this model.

3.3. Applicative Case

In order to complete our study, we propose to analyze the performance of the different ODE solvers with our simulation platform RHENOMS, which models a glutamatergic synapse by integrating a large number of models of receptors, transporters, enzymes and diffusion models. This simulation platform is schematically represented on Figure 4, which depicts the different elements that are integrated. The schematic does not depict the real complexity of the actual presynaptic compartment, which includes models of action potential, calcium diffusion, neurotransmitter release and diffusion. Similarly, models of cytosolic calcium, Long Term Potentiation (LTP) and postsynaptic potential are not represented in the postsynaptic compartment, nor are models of extrasynaptic Acetylcholinesterase (AChE) and Glutamate Transporters (GluT). In short, the glutamatergic synapse platform integrates more than 300 equations to be solved at each integration step, 144 variable states for 21 bilinear synaptic receptor kinetic models.

Computing the glutamatergic RHENOMS synapse with TR-BDF2 algorithm with *Tol2* is beyond the capacity of the computer used for this study, and should be performed on a computer cluster with much more memory. In fact, the second set of tolerance places TR-BDF2 in its worst configuration. TR-BDF2 algorithm does not need a restrictive tolerance to provide accurate enough results. In addition, users do not change the default tolerance parameters, which are commonly set close to our first set (*Tol1*). For these reasons, we present algorithm performances with the two stimulation protocols (P1 and P2) and the first tolerance set only (*Tol1*).

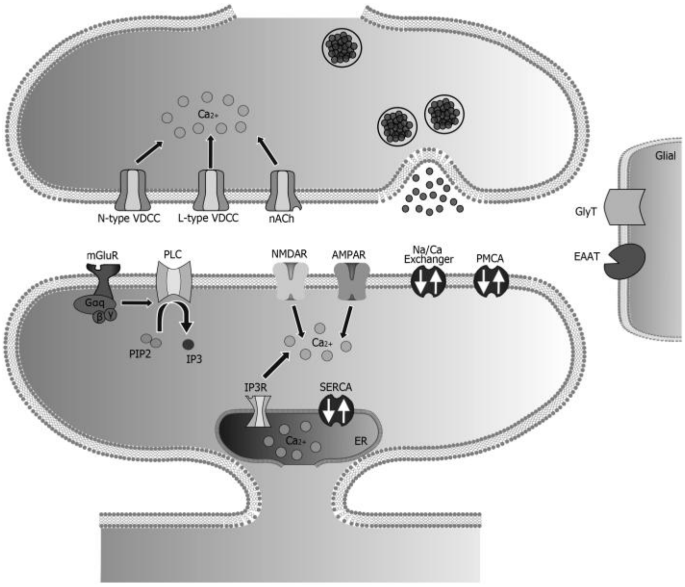


Figure 4. Schematic representation of RHENOMS glutamatergic synapse simulation platform.

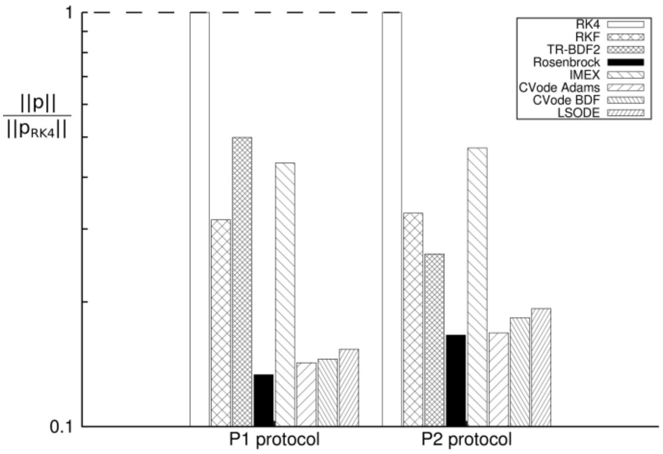


Figure 5. Normalized norm values for the 8 solver algorithms for the glutamate synapse platform stimulated with P1 and P2. All $||p||$ norms are normalized with the reference RK4 $||p_{RK4}||$ norm.

As shown in Figure 5, the Rosenbrock solver is the solver that provided the smallest norm for both protocols. In contrast, IMEX, which is a hybrid solver with a fourth-order RK scheme, generate a large norm value. Surprisingly, TR-BDF2 algorithm improves it performance between P1 and P2 protocol. This could come to his step-size adaptation t_{stop} [5] with which TR-BDF2 readapt the integration step-size around a brutal variation (i.e. a stimulation input). In theory, more the model will receive stimulation and more the algorithm will be performer. It is the sole algorithm with a better performance with P2 as compared to P1.

The glutamate synapse platform is clearly very complex and it is very hard to determine its stiffness or its non-linearity degree. Nevertheless, we are able to identify which solver was optimal to use for this simulation platform, which is the hybrid Rosenbrock solver.

4 Conclusions

Computational neuroscience encompasses a wide range of kinetic models with very different characteristics. It is very difficult to select the appropriate algorithm to solve the ODE systems representing biological models. Indeed, in order to select the right algorithm, as suggested by Rice, users need to know the model, the simulation tool and the ODE solver structure. A benchmark comparing various ODE solver algorithms or a recommendation could help users to select the most appropriate algorithm for a given simulation. We benchmarked 8 ODE solvers using two types of kinetic models with two stimulation protocols. This benchmark is done in Java, and it is important to note that the conclusions would not change if we change the programming language, while respecting good coding practice. Conclusions are summarized in Table 1.

Model	**Protocol**	**Tolerance**	**Best ||*p*||**	**Less ||*p*||**
Slow model: NMDA receptor	P1	Tol1 Tol2	IMEX IMEX	RKF TR-BDF2
	P2	Tol1 Tol2	IMEX IMEX	RKF TR-BDF2
Fast model: N-type VDCC	P1	Tol1 Tol2	LSODE LSODE	TR-BDF2 TR-BDF2
	P2	Tol1 Tol2	LSODE LSODE	TR-BDF2 TR-BDF2
Glutamatergic synapse	P1	Tol1	Rosenbrock	TR-BDF2
	P2	Tol1	Rosenbrock	IMEX

Table 1. Summary of best and less (highest) performing solvers for the two types of models with two different stimulation protocols, and the applicative case. P1 stands for the single event protocol and P2 stands for the repeated events protocol.

According to Table 1, TR-BDF2 is the less appropriate algorithm for all the models we use in this study with the selected stimulation protocols and tolerance parameter sets. TR-BDF2 is an ODE solver combining a trapezoidal scheme followed by a second-order Backward Differential Formula (BDF2), and, as shown in our results, a relative tolerance set below or equal to $1e^{-6}$ reduced this algorithms performance. In Fact, the restrictive tolerance sets produced the worst overall ||*p*|| norm performance value on TR-BDF2 solver. This algorithm could produce better performances and accurate enough results with a large tolerance, but a more economical (in terms of performance) solver could give accurate enough results as well.

IMEX algorithm is the best choice for a slow model, such as the NMDA receptor. LSODE is the best choice for a fast model, such as a Voltage-Dependent Calcium Channel and Rosenbrock solver is the more appropriate for the glutamatergic synapse model, which is considered as a complex model.

However, most of the time, users have not a priori information on the models or tools structure. In that case, we recommend LSODE solver. This solver is neither the best nor the worst, but because of its ability to switch between an explicit and an implicit scheme, our results indicate that it gives stable performance for all models, regardless of the stimulation protocol. In fact, if we change the stimulation protocol for any model, LSODE solver's performances does not change significantly.

As a comparison with others simulation platforms or softwares, the *Neuron* software uses CVODE algorithm with a manual selection between the included two scheme. The *Copasi* tool (biochemical network simulator) uses a fourth-order hybrid Runge-Kutta algorithm in addition to LSODE algorithm which integers Petzolds stiffness detection. *BioUML* software developers have made the same choice as Copasi. They use to a fourth-order hybrid Runge-Kutta algorithm in addition to CVODE algorithm implemented in Java language. *Matlab* follows our same decision which consists to integer several resolution algorithms and to allows users to choice their algorithm based on simulated model.

It is in general very difficult to determine the stiffness degree of a particular kinetic model, as this requires a careful analysis of the set of differential equations. In the benchmark method presented, we did not determine the stiffness degree of a model. However, this piece of information is not necessary to perform this study. We are able to make a decision regarding the choice of solver without any a priori information on stiffness, model equations or solver structure. This solver comparison and the resulting recommendations could therefore facilitate the choice of solvers for simulation of kinetic receptor synaptic models for any users.

A possible extension of this work would be to generate an adaptive algorithm, which could select, from these 8 ODE solvers, the best algorithm during the simulation, depending on model properties, input protocol and algorithm performance. This would result in a new solver, which could generate an accurate enough solution with maximal performance.

Acknowledgements

We would like to thank Lhassane Idoumghar at the University of Haute Alsace (France) and Tameem Albash at the University of Southern California (USA, CA) for constructive discussions, and the operational team of Rhenovia Pharma who helped for the elaboration of this study. Rhenovia obtained financing by the French National Agency of Research and Technology (ANRT) with a CIFRE scholarship.

References

[1] J. R. Rice, The algorithm selection problem. *Advances in Computer*, pp. 65–118, 1976.

[2] J-M Bouteiller, M. Baudry, S. Allam, R. Greget and S. Bischoff, Modeling glutamatergic synapses: Insights into mechanisms regulating synaptic efficacy. *J. Integr Neurosci*, pp. 185-197, 2008.

[3] E. Hairer and G. Wanner, Runge-Kutta and extrapolation methods. Solving ordinary differential equations I: Nonstiff problems. *Springer*, Vol. 1, pp. 132, 2010.

[4] P. Bogacki and L. Shampine, An efficient Runge-Kutta (4,5) pair. *Computers & Mathematics with Applications*, pp.15-28, 1996.

[5] R. Bank, W. Couggrahm, W. Fichtner, E. Grosse and D. Rose, Transient simulation of silicon devices and circuits. *IEEE Transactions on electron devices*, pp. 1992-2007, 1985.

[6] L. Shampine, Implementation of Rosenbrock methods. *ACM Transactions on Mathematical Software*, pp. 98-113, 1982.

[7] K. Toshiyuki, IMEX Runge–Kutta schemes for reaction–diffusion equations. *Journal of Computational and Applied Mathematics*, pp. 182-195, 2008.

[8] S. Cohen, A. Hindmarsh, CVODE, a stiff/nonstiff ODE solver in C. *Computers in Physics*, pp. 138–143, 1996.

[9] A. Hindmarsh, ODEPACK a systematized collection of ODE solvers. *IMACS Transactions on Scientific Computation*, pp. 55-64, 1982.

[10] L. Petzold, Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM*, pp. 136–148, 1983.

[11] S. Schorge, S. Elenes and D. Colquhoun, Maximum likelihood fitting of single channel NMDA activity with a mechanism composed of independent dimers of subunits. *The Journal of Physiology*, pp. 395–418, 2005.

[12] A. Hodgkin and A. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, pp. 500–544, 1952.

[13] N. Ambert, R. Greget, O. Haeberlé, S. Bischoff, T. Berger, J-M Bouteillern and M. Baudry, Computational studies of NMDA receptors: differential effects of neuronal activity on efficacy of competitive and non-competitive antagonists. *Open Access Bioinformatics*, pp. 113–125, 2010.

[14] D. Jaffe, W. Ross, J. Lisman, N. Lasser-Ross and H. Miyakawa, A model for dendritic Ca2+ accumulation in hippocampal pyramidal neurons based on fluorescence imaging measurements. *Journal of Neurophysiology*, pp. 1065–1077, 1994.

[15] P. Poirazi, T. Brannon and B. Mel, Arithmetic of subthreshold synaptic summation in a model CA1 pyramidal cell. *Neuron*, pp. 977–987, 2003.

Toward a cyber Galaxy ?

Christophe CARON[1], Wilfrid CARRE[1], Alexandre CORMIER[1], Sandra DEROZIER[2], Franck GIACOMONI[3], Olivier INIZAN[4], Gildas LE CORGUILLE[1], Alban LERMINE[5], Sarah MAMAN[6], Pierre PERICARD[1] and Franck SAMSON[2]

[1] CNRS, UPMC, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France
{christophe.caron, wilfrid.carre, alexandre.cormier, gildas.lecorguille, pierre.pericard}@sb-roscoff.fr

[2] INRA, UR1077, MIGALE, Centre de Jouy-en-Josas, 78352, Jouy-en-Josas, France
{sandra.derozier, franck.samson}@jouy.inra.fr

[3] PFEM, UMR1019 INRA, Centre Clermont-Ferrand-Theix, 63122, Saint Genes Champanelle, France
franck.giacomoni@clermont.inra.fr

[4] INRA, UR1164,, Route de St Cyr, Versailles, France
olivier.inizan@versailles.inra.fr

[5] Institut Curie, INSERM, U900, Bioinformatics and Computational Systems Biology of Cancer, 75248 Paris, France
alban.lermine@curie.fr

[6] INRA, UMR444, Laboratoire de Génétique Cellulaire, Centre de Toulouse Auzeville, 24 Chemin de Bordé Rouge, 31320 Auzeville-Tolosane, France
sarah.maman@toulouse.inra.fr

Abstract *The success of the open web based platform "Galaxy" is growing among diverse scientific communities. The French Institute of Bioinformatics - IFB wish to initiate a collaborative work dedicated to scientific workflows and especially to the platform Galaxy. We report here the main items on which future collaborations could be build: (i) software and hardware architecture, (ii) tools integration and (iii) training.*

Keywords Galaxy, training, workflow, NGS, tools integration, data sharing

Vers une Cyber Galaxy ?

Résumé *Le portail Galaxy dédié à l'activité de bio-analyse connaît un succès croissant au sein de multiples communautés scientifiques. L'Institut Français de Bioinformatique (IFB) souhaite mener une action transversale dédiée aux workflows d'analyse de données et en particulier à la plateforme Galaxy. Nous présentons ici les axes majeurs de cette action en termes d'architecture logicielle et matérielle, d'intégration d'outils et de formations.*

Mots-clés Galaxy, formation, workflow, NGS, intégration d'outils, partage de données.

1 Introduction

L'Institut Français de Bioinformatique (IFB) a pour mission la coordination de la communauté bio-informatique nationale. Dans un contexte où l'analyse des données à haut-débit modifie considérablement la façon de mener des analyses avec la mobilisation de nouvelles infrastructures, de nouveaux outils et de nouvelles compétences, l'IFB a décidé de mener une action transversale autour des solutions dédiées aux « workflows » d'analyse de données, avec en particulier la plateforme Galaxy [1].

Galaxy est une plateforme scientifique web, libre et « open source », permettant la mise à disposition d'outils d'analyses orientés principalement pour la bioinformatique (NGS, Métabolomique, etc…) et les statistiques à un large panel d'utilisateurs. L'émergence de telles plateformes est liée au fait qu'une grande majorité des outils couramment utilisés, le sont via la ligne de commande, en limitant l'accès aux seuls spécialistes. Ainsi, Galaxy propose une méthodologie d'intégration d'outils issus de sources multiples et dispose d'un gestionnaire de workflows intuitif, d'un gestionnaire d'historique assurant la reproductibilité des analyses et d'un environnement de partage des données, des résultats, des outils, des worflows et des méthodologies d'analyses. L'un des objectifs du projet Galaxy est, via l'interfaçage web unifié d'outils et de fonctionnalité, de rendre accessible à des non-bioinformaticiens la réalisation d'analyses *in silico* évoluées.

Il existe autour de la plateforme scientifique Galaxy une communauté très active d'utilisateurs et de développeurs qui couplée aux fonctionnalités de transfert d'outils d'une instance d'un laboratoire à une autre fait que cet « environnement de travail » favorise le partage des développements et des collaborations entre les producteurs d'algorithme et les analystes.

Dans ce contexte, de nombreuses initiatives ont vu le jour depuis quelques mois autour de la plateforme Galaxy : communauté Galaxy-France, école bio-informatique Aviesan, Groupe Galaxy Aplibio. Afin de fédérer et structurer ces actions au niveau national, un groupe de travail a été constitué autour de plusieurs plateformes IFB (ABiMS, Curie, Genotoul/SIGENAE, MIGALE, URGI), et d'une plateforme de l'infrastructure nationale MetaboHub (PFEM). L'objet principal de cet article est d'exposer les axes de travail de cette action transversale en mettant l'accent sur les premières actions structurantes autour des bonnes pratiques de développements et de déploiement, ainsi que des formations qui verront le jour en 2013.

2 Axes de travail

Avec l'avènement des infrastructures réparties (localisation des données, multiplicité des outils, etc.), le déploiement des composants nécessaires au partage des données et aux traitements peut rapidement devenir une réelle difficulté, limitant ainsi les performances ou la qualité du dispositif. Avec la généralisation des infrastructures reposant sur Galaxy, il nous est apparu important de valoriser les retours d'expérience des plateformes qui ont été confrontées à différents verrous. Il s'agit donc pour le groupe de travail de proposer et partager des schémas organisationnels et techniques, en vue de proposer une infrastructure Galaxy cohérente avec la stratégie de l'IFB, notamment dans un contexte de Cloud académique.

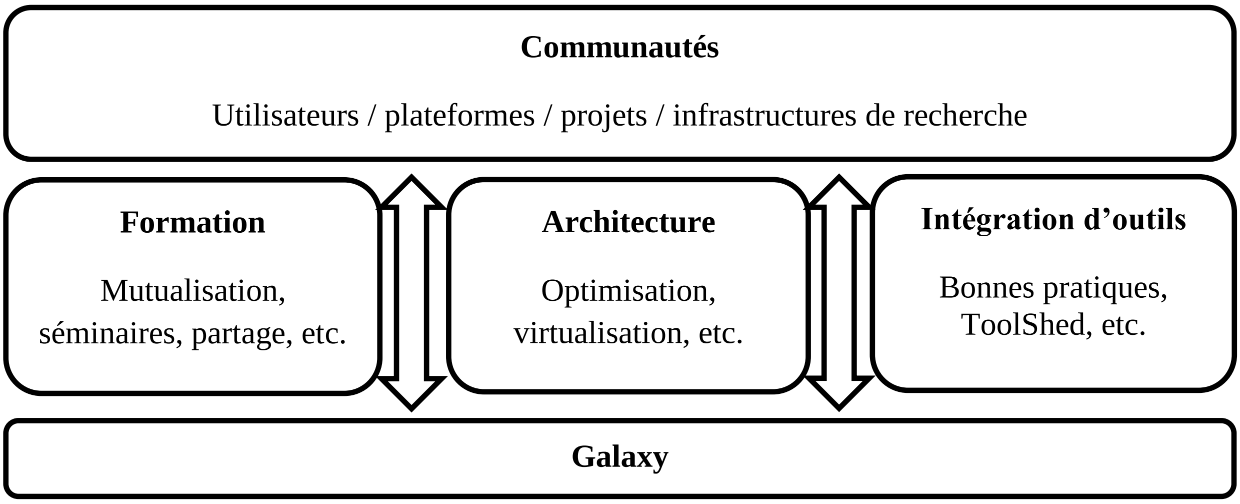

Communautés
Utilisateurs / plateformes / projets / infrastructures de recherche
Formation
Mutualisation, séminaires, partage, etc.
Architecture
Optimisation, virtualisation, etc.
Intégration d'outils
Bonnes pratiques, ToolShed, etc.
Galaxy

Figure 1. Axes autour du framework Galaxy travail

2.1 Architecture logicielle et matérielle

L'environnement Galaxy est constitué de plusieurs composants comme par exemple le gestionnaire de « jobs », le frontal web et la partie bases de données, dédiés à des tâches bien définies. Son déploiement et sa mise en œuvre peuvent présenter différentes finalités (formation, phases de prototypage de développement et d'exploitation, etc.), ce qui peut ainsi induire un niveau de qualité de service variable en termes de disponibilité et de scalabilité.

Ces différents contextes d'exécution font souvent appel à différentes techniques : virtualisation des instances, installation sur poste de travail, partage des données plus ou moins évolué, etc. Un des premiers résultats a consisté à valider la capacité de Galaxy à tenir la montée en charge tout au long d'une école thématique (Ecole Bio-informatique Aviesan, janvier 2013 - http://galaxy-ecole.sb-roscoff.fr/).

Les modifications et les optimisations apportées ont permis de supporter un grand nombre de connexions simultanées, dans un contexte d'hétérogénéité à plusieurs niveaux : typologie des analyses, volumétrie, diversité des codes parallèles, niveau de complexité des workflows, etc. Il a ainsi été possible de proposer un environnement supportant plusieurs dizaines d'utilisateurs simultanés sur des workflows d'analyses pouvant atteindre une centaine d'étapes. En continuant à explorer les différentes pistes d'optimisation, un des objectifs sera de proposer différents scénarios d'implémentation tenant compte du contexte d'usage, en puisant dans les retours d'expérience de la communauté.

Plus largement, les objectifs de ce volet consisteront à aborder les problématiques d'exploitation en environnement de production, les possibilités d'automatisation des tâches de déploiement et les solutions de monitoring d'instances de serveurs Galaxy.

2.2 Intégration d'outils

Les analyses de données à haut débit font appel à de nombreuses applications réparties sur un ou plusieurs sites. L'intégration de ces applications peut dès lors se révéler fastidieux. Ceci est d'autant plus vrai dans le cadre de déploiement de workflows, où s'enchaînent de nombreuses étapes et traitements. La plateforme Galaxy apporte des solutions concrètes à ces situations en centralisant tous les outils nécessaires à l'analyse et en proposant un modèle d'intégration simple pour ces outils [2].

L'objectif de cet axe est de proposer des outils et des processus facilitant l'interfaçage des outils dans une instance Galaxy. Nous pourrons aborder des situations comme (i) la montée en version d'outils, (ii) la maintenance d'outils « maison » en relation avec les montées en version de l'instance Galaxy, (iii) le test d'outils et de workflows nouvellement intégrés ou mis à jour.

Cette approche peut aussi bien passer par la rédaction d'un guide des bonnes pratiques, que par une veille technologique autour des méthodes proposées par la communauté.
Afin d'optimiser, en termes de facilité et de rapidité, l'intégration d'outils dans les instances locales de Galaxy, un des objectifs sera également de réfléchir sur l'intérêt du déploiement d'un « ToolShed » Galaxy propre à l'IFB prônant un haut niveau de qualité du descriptif et du contenu des outils partagés.

2.3 Diffusion, valorisation et formations

Les différentes plateformes du groupe de travail sont aujourd'hui amenées à proposer des formations Galaxy pour favoriser son apprentissage (Curie, Genotoul, ABiMS, MIGALE, URGI). Elles utilisent également Galaxy comme support de formation dans les principaux domaines de recherche en bio-informatique : analyses de données de transcriptome, détection de SNP, ChIP-seq, etc. Certaines de ces formations sont aussi proposées lors de programmes internationaux comme la conférence Galaxy 2013 à Oslo (ChIP-seq – Institut Curie). Ce mouvement d'appropriation de Galaxy par la communauté démontre une fois de plus son caractère essentiel comme outil dans le transfert de compétences. La plateforme Genotoul/Sigeneae propose également depuis plusieurs mois, des autoformations en ligne ainsi qu'une FAQ ouvertes à l'ensemble des utilisateurs. Ces modules d'autoformation (Initiation à Galaxy, etc.) assurent une prise en main du « workbench » Galaxy et une transmission des bonnes pratiques d'utilisation.

Au final, ces formations constituent un point d'entrée idéal pour les biologistes et les bio-analystes pour l'apprentissage des bonnes pratiques d'analyses. Elles sont aussi un moyen de démocratisation dans l'usage des concepts autour des workflows.

Un des objectifs du groupe sera, en coordination avec l'ensemble du dispositif IFB, de favoriser la mutualisation des actions de formations au niveau national, et d'assurer une transition vers de nouveaux usages comme le E-learning, en utilisant notamment les compétences acquises par le pôle toulousain.

Conclusion & Perspectives

Les premiers travaux présentés ont permis de fédérer une première communauté qui avait déjà initié un certain nombre de réflexions autour de Galaxy. Les prochains mois vont permettre d'avancer sur les différents axes présentés, avec notamment la mise en ligne d'un site documentaire.

Il s'agira aussi de lier cette action d'animation avec les autres infrastructures nationales de recherche (MetaboHub, EMBRC-France, etc.) financées par le programme ANR des Investissements d'Avenir. L'utilisation de cette action structurante autour de Galaxy pourrait être une première étape vers un partage plus important des usages et bonnes pratiques en termes de développement, mais aussi vers la création de passerelles entre ces différents projets. Des collaborations scientifiques actuelles illustrent de belle manière que l'environnement « Galaxy » peut devenir un excellent média et une passerelle entre les communautés que sont les bio-informaticiens et les biologistes.

Enfin, afin de présenter les résultats des développements en cours, mais aussi d'enrichir le partage de connaissances, nous organiserons un premier séminaire ouvert à la communauté IFB à l'automne. Ce séminaire de deux jours proposera une session avec des retours d'expérience Galaxy, et une formation aux bonnes pratiques (intégration d'outils, E-learning, etc.). Cette animation a pour vocation d'être élargie à d'autres entités (plateformes, communautés, etc.) dans le cadre de projets collaboratifs soutenus par l'IFB.

Acknowledgements

This work was supported by ANR programs 'Investissements d'Avenir': MetaboHub, EMBRC-France, France Génomique, 'Institut Français de Bioinformatique'.

References

[1] B. Giardine, C.Riemer, R.C.Hardison, R.Burhans, L.Elnitski, P.Shah, Y.Zhang, D.Blankenberg, I.Albert, J.Taylor, W.Miller, W.J.Kent and A.Nekrutenko, Galaxy: A platform for interactive large-scale genome analysis. Nucleic Acids Res., 15: 1451-1455, 2005.

[2] J.Goecks, A .Nekrutenko, J.Taylor and The Galaxy Team, Galaxy : a comprehensive approach for supporting accessible, reproducible, and transparent computaional research in the life sciences. Genome Biology, 11:R86, 2010.

Mapsembler2, assemblage ciblé rapide et visualisation de graphe d'assemblage

Alexan ANDRIEUX[1], Charles DELTEL[1], Rayan CHIKHI[2] and Pierre PETERLONGO[1]

[1] INRIA Rennes, Bretagne-Atlantique, EPI GenScale, Rennes, France
{alexan.andrieux, charles.deltel, pierre.peterlongo}@inria.fr
[2] Department of Computer Science and Engineering, The Pennsylvania State University, USA.
chikhi@psu.edu

Mots clefs Assemblage ciblé, Graphes de de Bruijn, Visualisation, Séquençage haut-débit

1 Introduction et contexte

Il est aujourd'hui relativement aisé de séquencer à moindre coût des échantillons biologiques. La difficulté réside dans l'extraction de connaissances une fois les données séquencées. En l'absence de génome de référence, assembler le matériel séquencé nécessite beaucoup de ressources de calcul (CPU, RAM). De plus, l'assemblage peut être le résultat d'un choix heuristique, effectué lorsque plusieurs possibilités d'assemblage coexistent ce qui conduit à une perte de la diversité dans les régions répétées des génomes. Les assemblages produits peuvent donc être incomplets et/ou chimériques [1,2]. C'est une difficulté majeure dans le cas de données complexes à assembler, tels que les génomes polyploïdes ou les méta-génomes.

Il apparait cependant que dans de nombreuses situations, il n'est pas indispensable de reconstruire une séquence de référence complète pour apporter de nouvelles connaissances biologiques. C'est le cas lorsque que l'on possède un *a priori* sur un fragment de séquence connu, et que l'on recherche son "contexte" dans des données de séquençage non assemblés.

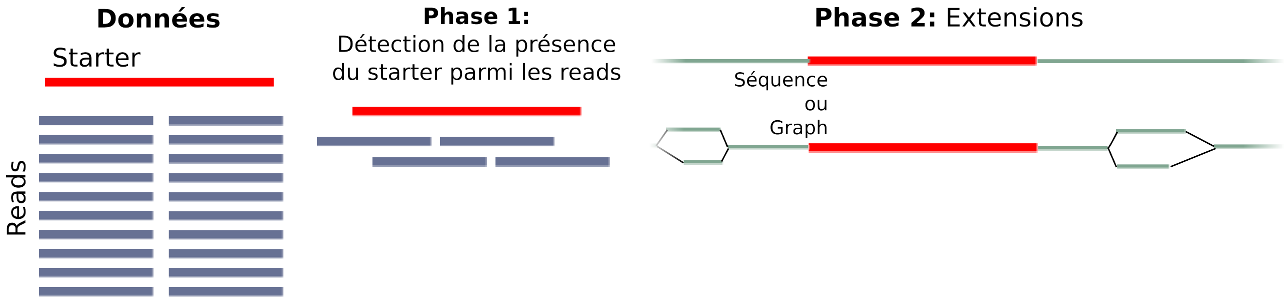


Figure 1. Aperçu des deux phases de l'approche MAPSEMBLER

Dans cette optique, une première version d'un outil, nommé MAPSEMBLER, a récemment été proposée [3]. Dans le paragraphe suivant, nous décrivons succinctement le fonctionnement de MAPSEMBLER. La suite de cet article est dédiée à la description d'améliorations récemment apportées.

MAPSEMBLER nécessite un ou plusieurs jeux de données de séquençages NGS (les *reads*) et un ou plusieurs fragments de séquences appelés *starters*. Premièrement, MAPSEMBLER détecte la présence de chaque starter parmi les reads (Fig. 1, phase 1). Plus précisément, un starter est considéré présent dans les reads si il est possible de réaliser un assemblage des reads contenant une séquence proche du starter (à un nombre maximal de substitutions près). Deuxièmement, pour chaque possibilité d'assemblage d'un starter, MAPSEMBLER construit son contexte (Fig. 1, phase 2). En fonction des paramètres demandés, le contexte construit par MAPSEMBLER est soit (i) une séquence nucléique autour de chaque starter, ou (ii) un graphe de séquences pour chaque starter, tel qu'il existe un chemin entre chaque nœud et le starter. Dans le cas (i), la sortie est un fichier au format *FASTA* ou *FASTQ* (selon le type de fichier d'entrée), et dans le cas (ii), la sortie est dans un format standard de graphe *json* (`http://www.json.org/`).

2 Nouveautés de MAPSEMBLER2 vis à vis de MAPSEMBLER

Nous proposons aujourd'hui une nouvelle version de MAPSEMBLER qui apporte des améliorations en terme de performances, en terme de type de sorties et en terme de visualisation des résultats graphiques.

2.1 Accélération des calculs

L'un des points clefs de MAPSEMBLER est une faible consommation de mémoire vive, afin d'être utilisable sur un ordinateur de bureau classique. Le principe de MAPSEMBLER est de réaliser des assemblages ciblés itératifs, une sous-partie différente des reads étant indexée à chaque itération. Cette approche occupe une empreinte mémoire minime (quelques méga-octets), mais nécessite un temps d'exécution long (plusieurs heures, selon la taille des entrées et la taille de l'assemblage souhaité), car les reads sont parcourus de nombreuses fois.

Récemment, Chikhi et Rizk [4] ont proposé une structure de données permettant d'assembler entièrement un génome avec une faible empreinte mémoire (un génome humain est assemblé en moins de 6 GB de RAM). Leur approche est basée sur une représentation du graphe de de Bruijn à l'aide d'un filtre de Bloom, ainsi qu'une autre structure gommant l'aspect probabiliste du filtre de Bloom. Cette approche a été implémentée dans MAPSEMBLER2, et a été adaptée à l'assemblage ciblé. L'utilisation mémoire reste compatible avec une utilisation du logiciel sur un ordinateur de bureau. L'absence de parcours multiple des reads accélère drastiquement les temps de calcul par rapport à la version précédente.

		MAPSEMBLER	MAPSEMBLER2	
Temps	Phase1	1m10	1m10	
	Phase2 : création de l'index	∅	2m46	
	Phase2 : extensions	1h18m49m	1m05	
Temps	Total	1h19m59m	Avec creation de l'index	6m01
			Sans création de l'index	2m15
Mémoire		70MB	Avec creation de l'index	4GB
			Sans création de l'index	96MB

Table 1. Comparaison temps et mémoire entre MAPSEMBLER et MAPSEMBLER2.

Afin de mettre en évidence ces différences, nous avons appliqué MAPSEMBLER et MAPSEMBLER2 sur un jeux de donnée Illumina d'orang-outan (SRA011022), composé de 30.9 millions de reads. Parmi ces reads, nous en avons sélectionné trois (de manière aléatoire) comme starters. Les résultats en terme de temps de calcul et de mémoire consommée par MAPSEMBLER et MAPSEMBLER2 sont présentés Table 1. En outre nous avons également différencié le cas où MAPSEMBLER2 est utilisé sur des données pour lesquels l'index a déjà été créé une fois et n'est pas reconstruit, évitant le temps de création de la structure d'index, stockée sur disque. Ce cas se présente si un nouveau *starter* doit être testé pour des données sur lesquels un ou plusieurs *starters* avaient déjà été testés. D'une manière générale, nous constatons que les temps de calculs ont étés très largement réduits dans MAPSEMBLER2 : par un facteur 16 lorsque l'index n'est pas créé et par un facteur 34 lorsque l'index a déjà été créé. L'empreinte mémoire de MAPSEMBLER2 est bien entendu plus importante que celle de MAPSEMBLER (4GB au lien de 70MB). Cependant, celle-ci reste acceptable pour une utilisation sur une station de travail actuelle. Notons en outre qu'il est possible de réduire l'empreinte mémoire de MAPSEMBLER2 au prix d'un temps de création de l'index plus long.

2.2 Choix du type d'assemblage ciblé

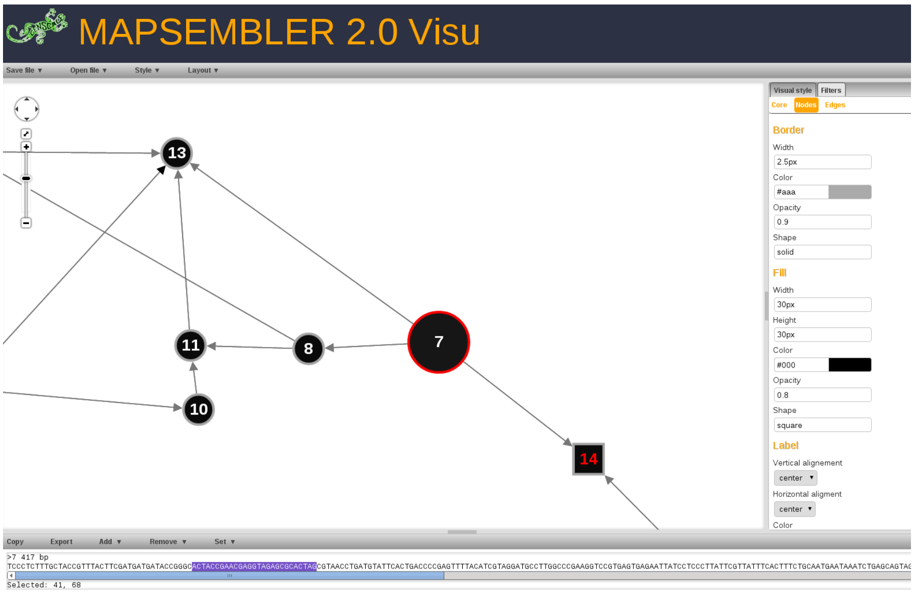
Le changement de structure de données, décrit dans la section précédente, permet plus de flexibilité dans le type d'assemblage ciblé. Les séquences assemblées peuvent désormais être des contigs ou unitigs [1]. Dans MAPSEMBLER2, l'utilisateur peut désormais choisir entre quatre types d'assemblages ciblés : séquences simples étants des *contigs* ou des *unitigs*, ou graphe dont les nœuds sont des *contigs* ou des *unitigs*. Ces différents types d'assemblage n'étaient pas possibles dans la version précédente.

[1] Un *unitig* résulte de l'assemblage sans qu'aucun choix n'ait été fait, en opposition aux *contigs*, qui sont également des résultats d'assemblages, mais pour lesquels les "petites" variations (SNPs, indels) ont été cachées.

En outre, nous pouvons noter, que dans la version précédente, par limitation liée au temps de calcul, un assemblage ciblé était nécessairement court : quelques centaines de nucléotides de chaque coté du starter. Pour certaines applications (e.g. un assemblage classique localisé, reconstruction d'évènements d'épissage alternatifs, . . .), il est souhaitable de réaliser un assemblage contenant des séquences plus longues, ce qui est désormais accessible avec MAPSEMBLER2.

2.3 Interface web de visualisation

Comme nous l'avons évoqué, une sortie possible de MAPSEMBLER2 est un graphe, au format *json*. Nous avons choisi ce format de graphe car, étant plus condensé, il est une bonne alternative aux formats équivalents (*xgmml, graphml*). Afin de permettre une exploitation aisée des fichiers de graphe produits par MAPSEMBLER2, nous proposons une interface de visualisation de ces résultats.



Border
Width: 2.5px
Color: #aaa
Opacity: 0.9
Shape: solid

Fill
Width: 30px
Height: 30px
Color: #000
Opacity: 0.8
Shape: square

Label
Vertical alignement: center ▾
Horizontal aligment: center ▾
Color

```
>7 417 bp
TCCCTCTTTGCTACCGTTTACTTCGATGATGATACCGGGCACTACCGAACGAGGTAGAGCGCACTAGCGTAACCTGATGTATTCACTGACCCCGAGTTTTACATCGTAGGATGCCTTGGCCCGAAGGTCCGTGAGTGAGAATTATCCTCCCTTATTCGTTATTTCACTTTCTGCAATGAATAAATCTGAGCAGTAG
Selected: 41, 68
```

Figure 2. Selection d'un nœud et affichage de la séquence qu'il contient.

Cette interface de visualisation à été développée en JavaScript/JQuery en se basant sur la librairie *Cytoscape.js* qui reprend les grandes idées de la version Java de *Cytoscape* [5]. Le choix d'une interface orientée web est justifié par la facilité de proposer aux utilisateurs une exploitation simple des données de sortie de MAPSEMBLER2 aussi bien de manière locale que distante. Aucune installation de logiciel tiers n'est nécessaire pour utiliser le visualiseur, un navigateur web usuel (Firefox, Chrome, Safari) étant suffisant.

L'interface que nous proposons permet à l'utilisateur de naviguer dans le graphe (zoom et déplacement), et de sélectionner un ou plusieurs éléments (Fig. 2) afin d'en visualiser les informations (séquence, couverture, qualité). L'utilisateur peut également appliquer une apparence globale au graphe (thème visuel appliqué à l'ensemble des nœuds et des arrêtes sans tenir compte de leurs propriétés) ou il peut appliquer une apparence différentielle à chaque élément du graphe (nœuds et arêtes) en fonction des informations contenues dans ces éléments. Ainsi, divers paramètres visuels (e.g. forme, couleur, taille, transparence) des éléments peuvent être affectés à des paramètres biologiques (longueur de séquence, couverture, qualité) (Fig. 3).

L'interface propose également plusieurs formes de mise en page (*layout*) permettant d'organiser le positionnement des nœuds en fonction des préférences de l'utilisateur. Un système de sauvegarde et de chargement de session est également proposé.

3 Conclusions

Nous proposons MAPSEMBLER2, une nouvelle version de l'outil MAPSEMBLER, permettant de réaliser des assemblages ciblés en quelques minutes. Le temps d'exécution de MAPSEMBLER2 est d'un ordre de magnitude

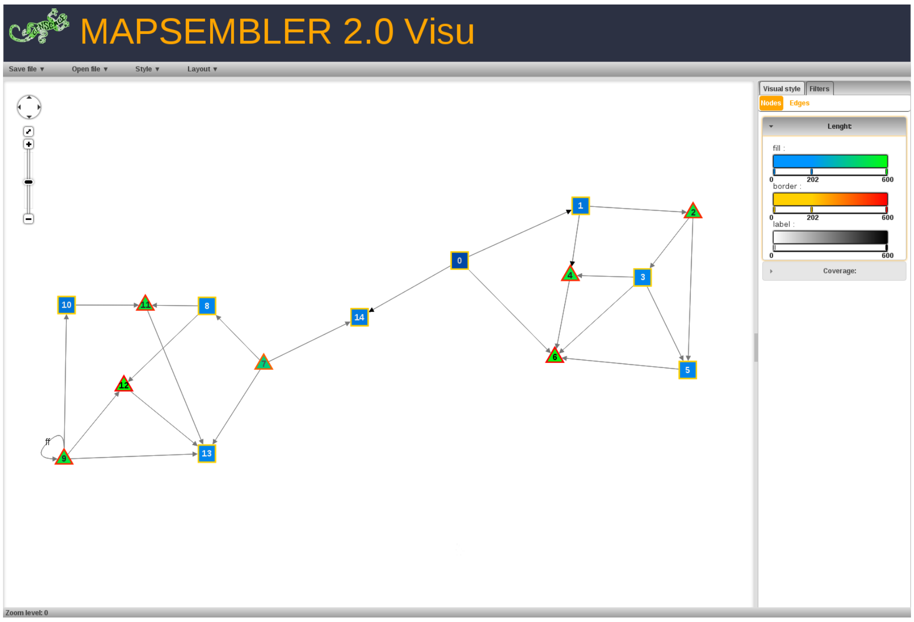


Figure 3. Exemple d'application de paramètres visuels en fonction de la longueur de séquence contenue dans chaque nœud : carré bleu à bord jaune de 1 bp à 200 bp et triangle vert à bord rouge de 201 bp à 600 bp.

inférieur à la version précédente, pour une qualité de résultats strictement équivalente. De plus, nous proposons une interface de visualisation *ad-hoc*, utilisable via un simple navigateur web. MAPSEMBLER2, sous licence CeCILL est librement téléchargeable : `http://colibread.inria.fr/`.

Remerciements

Ce travail a bénéficié du support d'une ADT (Action de Développement Technologique) Inria. Ce travail a reçu les support financier de l'ANR-12-BS02-0008 (Colib'read).

Références

[1] Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C. J., & Deng, H.-W. (2011). Comparative Studies of de novo Assembly Tools for Next-generation Sequencing Technologies. Bioinformatics (Oxford, England), 27(15), 2031–2037. doi :10.1093/bioinformatics/btr319

[2] Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. Nat Meth, 8(1), 61–65. Retrieved from http ://dx.doi.org/10.1038/nmeth.1527

[3] Peterlongo, P., & Chikhi, R. (2012). Mapsembler, targeted and micro assembly of large NGS datasets on a desktop computer. BMC Bioinformatics, 13(1), 48. doi :10.1186/1471-2105-13-48

[4] Chikhi, R., & Rizk, G. (2012). Space-efficient and exact de Bruijn graph representation based on a Bloom filter. In Lecture Notes in Computer Science (Ed.), WABI (Vol. 7534, pp. 236–248).

[5] Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., & Ideker, T. (2011). Cytoscape 2.8 : new features for data integration and network visualization. Bioinformatics (Oxford, England), 27(3), 431–2. doi :10.1093/bioinformatics/btq675

Resilience in a cheese ecosystem

Frédéric FER[1,2,3,4], Julie AUBERT[3,4] and Jean-Marie BECKERICH[1,2]

[1] AgroParisTech, UMR 1319 MICALIS, F-78350 Thiverval Grignon, France
[2] INRA, UMR 1319 MICALIS, F-78350 Thiverval Grignon, France
{frederic.fer, jean-marie.beckerich}@grignon.inra.fr
[3] AgroParisTech, UMR 518 MIA, F-75005 Paris, France
[4] INRA, UMR 518 MIA, F-75005 Paris, France
{frederic.fer, julie.aubert}@agroparistech.fr

Abstract *The resilience is the capacity for an ecosystem to return to a steady or cyclic state following a perturbation. To study this phenomenon, we have developed a RNA-seq based approach on a model cheese reduced ecosystem (Munster type) composed of nine species with sequenced and annotated genomes. The aim of this work is to develop methods to integrate, analyze and link diverse data (microbial growth, biochemical, physico-chemical and genomic data) in order to understand the microbial ecosystem resilience. This will provide to the scientific community a set of reusable and flexible tools to study with high-throughput methods microbial ecosystems.*

Keywords RNAseq, microbial ecosystems, metatranscriptomic analysis, proteolysis, deterministic model, differential expression

Résilience dans un écosystème fromager

Résumé *La résilience est la capacité d'un écosystème à maintenir ou rétablir une fonction d'intérêt malgré une perturbation. Pour étudier ce phénomène, nous avons mis en place une approche de type RNA-seq sur un écosystème modèle composé de 9 espèces connues et maîtrisées, l'écosystème fromager type Munster, sur lequel nous étudions une fonction majeure : la dégradation des caséines. Le but de ce travail est de développer des méthodes d'exploration type RNA-seq sur des écosystèmes microbiens, de façon à y étudier les mécanismes de la résilience. Cela permettra de fournir à la communauté scientifique un ensemble d'outils souples et réutilisables pour initier des études à haut débit sur les écosystèmes microbiens.*

Mots-clés Séquençage ARN, écosystèmes microbiens,analyse du métatranscriptome,protéolyse, modèle deterministe, expression différentielle

1 Introduction

1.1 Écosystèmes microbiens

Un écosystème microbien désigne un ensemble de micro-organismes cohabitant dans un même environnement. Cet environnement peut-être naturel (sol, océans, colon) ou artificiel comme dans certains processus agroalimentaires ou industriels (fermentation, épuration). Ces systèmes sont composés de multiples espèces, souvent mal connues et non cultivables, qui à travers leurs interactions permettent l'émergence de fonctions complexes. Les écosystèmes microbiens sont donc sources de nombreux ≪ services écologiques ≫ tant sur le plan de l'environnement que sur celui de la santé (digestion) de l'industrie ou de l'alimentation. La compréhension et la maîtrise de ces écosystèmes représentent donc aujourd'hui un énorme enjeu scientifique, sanitaire et économique.

1.2 Résilience

Une des fonctions complexes d'un écosystème est la résilience, terme designant la capacité d'un écosystème à maintenir ou rétablir une fonction d'intérêt malgré une perturbation. Cette perturbation peut être une modification de l'environnement ou de l'équilibre entre les espèces. Dans le cas simple de la disparition d'une espèce portant majoritairement une fonction d'intérêt, le maintien de cette fonction peut s'expliquer par deux phénomènes non exclusifs qui permettent de maintenir la fonction d'intérêt. Soit l'une ou plusieurs espèces participant également à cette fonction (redondance fonctionnelle) peuvent voir leurs effectifs augmenter, soit une ou plusieurs de ces espèces peuvent modifier leurs métabolismes, par exemple afin d'augmenter une production enzymatique. Dans ce second cas une approche RNAseq permettra d'observer ces modifications métaboliques.

Afin d'étudier ce phénomène nous travaillons sur un écosystème modèle dont les espèces sont connues. Les phénomènes de résilience sont identifiés via le suivi par dosage de certaine fonctions majeures de l'écosystème. Dans cette étude nous nous focalisons sur une fonction particulière de l'écosystème : la protéolyse (dégradation des protéines). Le suivi par comptage des proportions d'espèces permet grâce à un modèle cinétique simple d'identifier celles intervenant dans le phénomène de résilience, et notamment celle dont l'intervention s'explique par une modification de leurs métabolismes. Pour ces espèces le suivi par séquençage haut débit de la concentration des transcrits du méta-transcriptome (ensemble des transcriptomes de l'écosystème) permet d'identifier les gènes majeurs impliqués dans ce phénomène.

2 Étude de la dégradation de la caséine (protéolyse)

2.1 Données

Notre modèle biologique ([3]) est un écosystème microbien fromager de type « Munster ». L'avantage de travailler sur cet écosystème modèle est que les neuf espèces le composant sont connues, cultivables et séquencées. L'écosystème est ensemencé, puis le développement de l'écosystème (affinage) se fait en milieu stérile afin d'éviter la contamination par des espèces non désirées. Les espèces en présence étant d'halotolérances variables, nous augmentons la salinité de la matrice afin de perturber leurs proportions dans l'écosystème (projet ANR ExECO). Une seconde approche plus directe dont les resultats sont en cours d'acquisition est de pratiquer une omission d'espèce (projet ECOSTAB) (Fig. 1). Nous étudions ensuite l'impact de cette modification sur le maintien des fonctions majeures de l'écosystème, notamment la protéolyse.

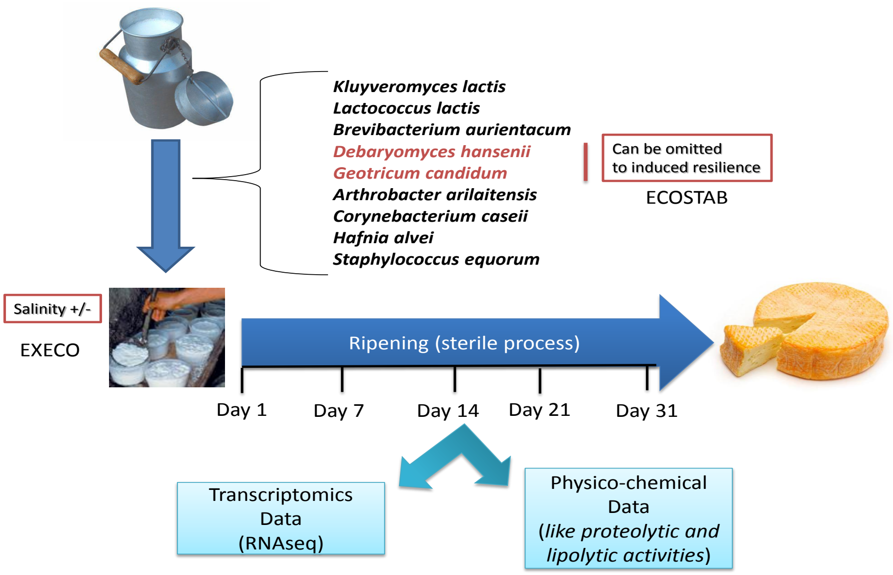

Kluyveromyces lactis
Lactococcus lactis
Brevibacterium aurientacum
Debaryomyces hansenii
Geotricum candidum
Arthrobacter arilaitensis
Corynebacterium caseii
Hafnia alvei
Staphylococcus equorum
Can be omitted to induced resilience
ECOSTAB
Salinity +/-
EXECO
Ripening (sterile process)
Day 1
Day 7
Day 14
Day 21
Day 31
Transcriptomics Data (RNAseq)
Physico-chemical Data (like proteolytic and lipolytic activities)

Figure 1. Protocole expérimental

Du fait de l'intérêt économique du fromage, les fonctions majeures de l'écosystème telle la protéolyse sont bien connues, elles sont suivies via le dosage de certains métabolites. Le comptage des espèces se fait sur boite. L'ARN est extrait directement de la matrice fromagère, et après une étape de transcription inverse puis de PCR le séquençage se fait sur un appareil SOLID (50 pb, single read). Les étapes de mapping puis de comptage s'effectuent respectivement via les logiciels Bowtie 0.12.7 et le pipeline Meteor (vers.06/12).

2.2 Modèle déterministe de la dégradation de la caséine

Pour identifier les espèces intervenant dans la protéolyse et dans son maintien nous utilisons un modèle de protéolyse basé sur des équations Michaeliennes et inspiré de celui de ([2].

2.2.1 Description du modèle : Lors de l'affinage du fromage, les caséines (80% des protéines du lait) sont dégradées en dipeptides via une enzyme menbranaire E_1, eux-même dégradés en acides aminés libres par une seconde enzyme E_2 cytoplasmique, relarguée dans le milieu lors des lyses cellulaires.

$$\textbf{[caseine]} \xrightarrow{[E_1]} \textbf{[dipeptides]} \xrightarrow{[E_2]} \textbf{[acides amines]}$$

Les enzymes sont produites par les neuf espèces en présence (de population $S_{i,i=1,\cdots,9}$). Le modèle suppose la stabilité au cours d'un affinage des taux de production journaliers des deux enzymes pour chaque espèce (noté respectivement α_1^i et α_2^i). Cela revient à dire que seules les variations des proportions des espèces expliquent les variations de la fonction protéolytique. En imposant cela, nous pourrons via l'étude des valeurs estimées de ces capacités protéolytique propres entre deux cinétiques, identifier les espèces pour lesquelles ces valeurs varient. Ces espèces sont en effet intéressantes car elles modifient leurs productions d'enzyme pour s'adapter à la perturbation, modification dont l'approche RNAseq permettra d'identifier les supports génétiques.

2.2.2 Équations : Selon le modèle, la vitesse de consommation de la caséine dépend de la concentration de l'enzyme E_1 et d'une constante réactionnelle $K_{caseines}$ spécifique de la réaction.

$$\frac{d[caseines]}{dt} = -E_1 \cdot \frac{[caseines]}{K_{caseines} + [caseines]} \tag{1}$$

L'enzyme E_1 est membranaire, donc sa concentration dans la matrice dépend des quantités des espèces en présence (chacune ayant son propre taux de production par cellule α_i^1) et de la dégradation spontanée dans la matrice (k_1 la constante de dégradation de la protéase).

$$\frac{dE_1}{dt} = \sum_i \left(\alpha_i^1 \frac{dS_i}{dt}\right) - k_1 E_1 \tag{2}$$

La vitesse de dégradation des dipeptides est fonction de la concentration de l'enzyme E_2^{extra} (constante réactionnelle $K_{dipeptides}$), les dipeptides sont un produit de la dégradation des caséines via E_1.

$$\frac{d[dipeptides]}{dt} = E_1 \cdot \frac{[caseines]}{K_{caseines} + [caseines]} - E_2^{extra} \cdot \frac{[dipeptides]}{K_{dipeptides} + [dipeptides]} \tag{3}$$

La concentration en dipeptidase (E_2) peut se séparer en une fraction cytoplasmique (E_2^{intra}) et une non cytoplasmique (E_2^{extra}). La dipeptidase cytoplasmique est produite lors de la synthèse de nouvelles cellules (au taux de production α_i^2 par cellule de l'espèce i) et disparait lors de la lyse cellulaire. La fraction présente dans la matrice E_2^{extra} libérée lors de la lyse est la seule qui intervienne dans la dégradation des dipeptides.

$$\frac{dE_2^{intra}}{dt} = \sum_i \left(\alpha_i^2 \frac{dS_i}{dt}\right) - k_l E_2^{intra} \tag{4}$$

$$\frac{dE_2^{total}}{dt} = \sum_i \left(\alpha_i^2 \frac{dS_i}{dt}\right) - k_l [E_2^{total} - E_2^{intra}] \tag{5}$$

La croissance d'une espèce de population S_i est modélisée par une fonction logistique prenant en compte chacun des nutriments $[N_n]$ consommés par l'espèce, via les paramètres de proportion de croissance de l'espèce expliquée par ce nutriment $\mu_{i,n}$ et le paramètre réactionnel propre $k_{i,n}$. Cette croissance est limitée par un taux de lyse k_l dans la matrice fromagère commun à toutes les espèces.

$$\frac{dS_i}{dt} = \sum_n \left(\mu_{i,n} \frac{[N_n]}{K_{i,n} + [N_n]} \right) S_i - k_l S_i \tag{6}$$

Les acides aminés libres de concentration $[aa]$ produits de la dégradation des dipeptides par E_2 participent à la croissance de certaines espèces.

$$\frac{d[aa]}{dt} = E_2^{extra} \cdot \frac{[dipeptides]}{K_{dipeptides} + [dipeptides]} - \sum_i \left(\mu_{i,aa} \frac{[aa]}{K_{i,aa} + [aa]} \right) S_i \tag{7}$$

2.2.3 Résultats : En se focalisant sur les deux premières équations, nous utilisons les données de comptage et les concentration en caséines et dipeptides pour estimer par une méthode de Newton-Raphson la valeur des capacités protéolytiques propres à chaque espèce (α_i^1), et ce pour chacune des cinétiques (salinité faible et forte). Les résultats (Fig. 3) montrent que l'espèce GC (*Geotrichum candidum*) semble la plus protéolytique, et que sa capacité propre estimée (α_{GC}^1) est identique en condition salée et moins salée. La seconde espèce la plus protéolytique est DH (*Debaryomyces hansenii*), et sa capacité propre est supérieure en condition plus salée.

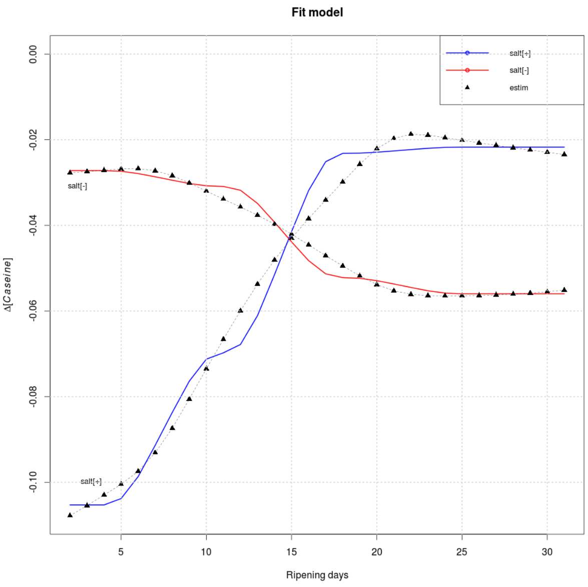


Figure 2. Ajustement du modèle de la consommation journalière de caséine (μg/ μL) aux données expérimentales. En trait plein les valeurs mesurées, les triangles représentent les valeurs recalculées à partir du modèle.

En condition salée GC est peu présente, DH au contraire se développe bien. Mais le modèle indique que l'augmentation de population de DH ne suffit pas à expliquer le maintien de la fonction protéolyse (α_{DH}^1 est plus elevé en condition salée). Nous pouvons faire l'hypothèse d'une modification métabolique de DH, augmentant sa capacité protéolytique propre en réponse à la baisse de population de GC afin de maintenir une fonction de protéolyse.

2.3 Données transcriptomiques

Une recherche de gènes différentiellement exprimés entre les deux conditions (salée et non salée) au jour 14 de l'affinage a été effectuée à l'aide du package edgeR ([4]) implémenté en R. Ce paquet intègre une normalisation TMM qui corrige notamment les différences de profondeur de séquençage ([1]). Le modèle d'analyse différentielle modèlise les données de comptage via une paramétrisation de la loi binomiale négative avec une méthode d'estimation de la surdispersion des données (par rapport à un modèle de Poisson). Les

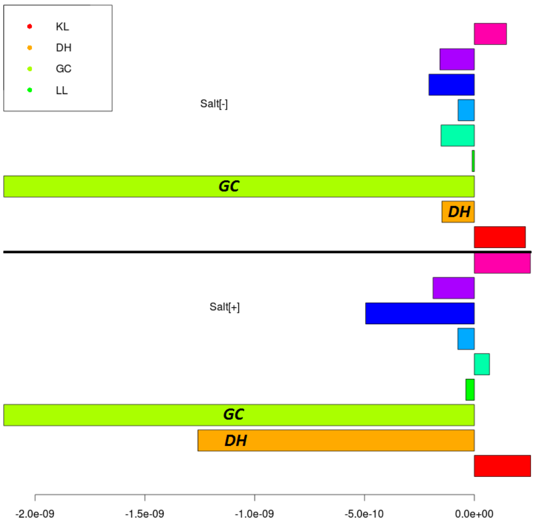


Figure 3. Pour chaque espèce, estimation des α_i^1 dans la condition salée, et non salée. On peut voir que les espèces GC et DH ont les valeurs estimées les plus fortes. Celle de DH varie entre les deux cinétiques.

gènes sont déclarés différentiellement exprimés de façon significatif après un ajustement des tests multiples par la procédure de Benjamini-Hochberg au seuil $\alpha = 5\%$. Nous avons ensuite séléctionné parmis les gènes différentiellement exprimés ceux appartenant à DH.

Cette analyse permet d'identifier une liste de gènes potentiellement impliqués chez DH dans le phénomène de maintien d'une activité protéolytique en absence de GC (Fig. 4). Nous mettons en place une méthode de régression afin d'identifier dans cette liste les gènes spécifiquement liés au phénomène de résilience dans la protéolyse.

3 Conclusion & perspectives

Notre modèle biologique est un écosystème parfaitement connu et maîtrisé qui nous permet d'étudier le phénomène de résilience dans les écosystèmes microbiens. Nous avons donc initié plusieurs cinétiques fromagères, avec une modification de la salinité du milieu, cela modifiant la proportion des espèces et donc les fonctions de l'écosystème

La construction d'un modèle cinétique de la dégradation des caseines nous a permis d'identifier les deux espèces majoritairement impliquées dans ce processus. Ce modèle montre que l'une des deux espèces (DH) semble modifier son métabolisme en réponse à la diminution en population de la seconde (DH) en milieu salé. Une analyse différentielle nous a permis d'identifier une liste de gène candidats, la mise en place de méthode de régression des données métagénomiques nous permettra ensuite d'identifier ceux pouvant être impliqués dans le phénomène de résilience lié à la dégradation de la caséine.

Une amélioration de la méthode d'analyse des données RNAseq prenant en compte les modifications de proportion des espèces pourrait être envisagée si l'on s'intèresse plus particulièrement aux transcrits d'une espèce de l'écosystème.

D'autres d'affinages ont été faits, pour lesquelles la perturbation induisant le phénomène de résilience est l'absence d'une des deux levures les plus impliquées dans la protéolyse (GC,DH). Les échantillons sont en cours de séquençage. Ces données nous permettront de valider et de préciser les rôles de ces levures dans le processus de protéolyse, et notamment de vérifier que la modification du métabolisme de DH n'est pas juste une réponse à la salinité du milieu.

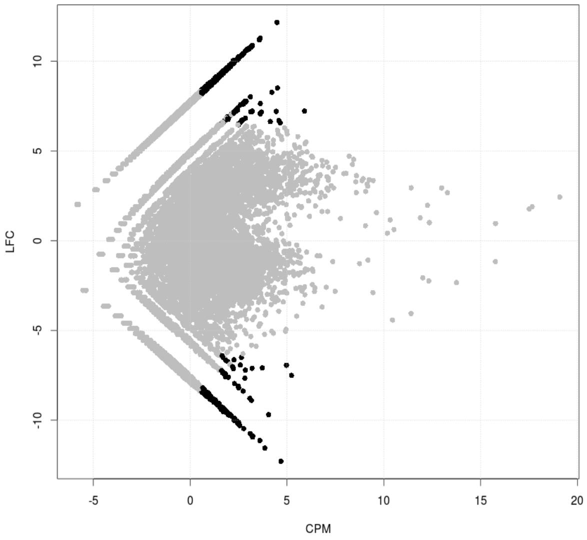


Figure 4. Résultat d'analyse différentielle entre les conditions salée (moyenne d'expression des gènes : μ_1) et non salée (μ_2). Chaque point représente un gène, représenté en fonction de son LFC(log($\frac{\mu_1}{\mu_2}$) et CPM(log($\frac{\mu_1+\mu_2}{2}$). En noir les gènes différentiellement exprimés au risque $\alpha = 0.5\%$.

Remerciements

Nous remercions l'ANR pour avoir permis la réalisation du projet ExECO, ainsi que tous les acteurs de ce projet pour nous avoir permis d'utiliser ces données et nous avoir fourni l'expertise nécessaire à une bonne compréhension des enjeux biologiques. Enfin nous remercions le métaprograme INRA MEM qui finance la thèse de F.Fer et la partie expérimentale ECOSTAB.

Références

[1] MA Dillies, A Rau, J Aubert, C Hennequet-Antier, M Jeanmougin, N Servant, C Keime, G Marot, D Castel, J Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 74(21) :6505–6512, 2012.

[2] J.K Kim, M Starzak, G.W Preckshot, R Marshall, and R.K Bajpai. Critical reactions in ripening of cheeses. *Applied biochemistry and biotechnology*, 45(1) :51–68, 1994.

[3] J. Mounier, C. Monnet, T. Vallaeys, R. Arditi, A.S. Sarthou, A. Hélias, and F. Irlinger. Microbial interactions within a cheese microbial community. *Applied and environmental microbiology*, 74(1) :172–181, 2008.

[4] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1) :139–40, January 2010.

MagSimus: Modeling Ancestral Genomes by SIMUlationS

Realistic modeling of gene order evolution in vertebrate genomes

Joseph LUCAS[1], Matthieu MUFFATO[1,2] and Hugues ROEST CROLLIUS[1]
[1] DYOGEN Lab, UMR8197 CNRS, 46 rue d'Ulm, IBENS, 75005, Paris, France
{jlucas, hrc}@ens.fr
[2] new address : EBI, Welcome Trust Genome Campus, Cambridge, BC10 1SD, Hinxton, United Kingdom
muffato@ebi.ac.uk

Abstract We present here a model of known evolutionary processes affecting gene order and its implementation into a program called MagSimus. MagSimus handles a range of known evolutionary events, such as branch specific chromosomal rearrangements (inversions, translocations, fissions and fusions) and gene events (apparitions, duplications and deletions). MagSimus can simulate a simple gene evolutionary history and a corresponding gene order evolution. However, such simple gene histories never approach real and complex genome wide gene evolution. MagSimus is thus also able to simulate gene order evolution given a forest of gene trees, i.e. a set of gene evolutionary trees covering the entire ancestral genome and specifying the ensuing history of every gene. To be effective, such an alternative approach requires the clustering of genes in the initial ancestral genome according to the similarity of their 'future' evolutionary histories. Since genes events are included in gene forest, the only remaining parameters concern chromosomal rearrangements. MagSimus is applied to the simulation of 5 extant genomes and their corresponding gene forest. Parameters are optimized by maximizing the similarities between the real and the simulated extant genomes. The results are theoretical chromosomal rearrangement rates and distribution of lengths of collinear gene orders (syntenic blocks). Remarkably, the theoretical values are very similar to known rates from the literature, thus suggesting that MagSimus generates realistic genome evolutionary histories.

Keywords Comparative genomics, simulation, syntenic blocks, chromosomic rearrangements, branch specific rates of evolution

Robust and complete pipeline for Infinium Methylation 450K data processing, normalization and bias correction.

Nizar Touleimat[1] and Jorg TOST[2]

[1] Laboratoire de Bioinformatique et d'Informatique
nizar.touleimat@cng.fr
[2] Laboratory for Epigenetics and Environment
Centre National de Génotypage, CEA - Institut de Génomique
Batiment G2, 2 rue Gaston Crémieux, CP5721, 91057 Evry Cedex, France
jorg.tost@cng.fr

Abstract

Infinium® Human Methylation 450K BeadChip (Illumina Inc., CA, USA) is an example of the huge progress that has been made in the development of array-based technologies for DNA methylation analysis. This array allows the simultaneous quantitative monitoring of more than 480,000 CpG positions and represents a good compromise between throughput, cost, resolution and accuracy for large-scale epigenotyping studies. However, there is currently no complete pipeline associated to this array, for quality control, elimination of unrelated signal variations, noise filtering and between-sample normalization. Moreover, we confirm the existence of an additional source of bias in 450k data, related to the two different assay chemistries combined on this array (Infinium I & Infinium II).

In this study we propose a complete preprocessing pipeline that addresses most current preprocessing issues related to Infinium® Human Methylation 450K BeadChip data. We also propose a new Infinium signal shift correction based on a subset quantile normalization (SQN) approach performed at the level of probe functional annotations. This approach uses Infinium I signal as 'anchors' to improve Infinium II methylation signal stability and accuracy, and to correct for the Infinium I/Infinium II shift.

We compared our preprocessing pipeline with existing approaches for the correction of the Infinium signal shift and evaluated all the approaches with Pyrosequencing data for a selection of CpG sites. Our SQN based approach outperforms other current correction approaches in terms of correction of the shift and quantitative estimation of the methylation status. In this study, we confirm the reproducibility and accuracy of Infinium Human Methylation 450K BeadChip data and we provide a complete pipeline that produces 'ready-to-analyze' and accurate methylation data.

Keywords Epigenetics, Methylome, Illumina Infinium methylation 450K, pipeline, preprocessing, normalization, bias correction, evaluation.

1 Introduction

We currently know that genetic variations are not sufficient to explain all inter-individual differences such as phenotypes (differences between twins for example), sensitivity to pathologies or response to treatments. Recent discoveries and developments in epigenetics provided new approaches to study relations between genome and phenotype. Epigenetics provides a new level of complexity where the genome is not simply considered as a "code" but also as an adaptable interface interacting with an individual's environment and behaviour. Knowledge on the multiple sides of the epigenome is therefore crucial for the understanding of the phenotypic variation of healthy and diseased cells, which is currently one of the main challenges in biomedical research. Molecular alterations of the epigenome, especially DNA methylation and miRNAs, have emerged as alternative targets of biomarker research and display potentially great clinical value. DNA methylation is the most-studied epigenetic variation and CpG dinucleotide methylation is involved in various biological processes and pathologies. Many studies have shown that complex diseases such as cancer [1,2,3] are associated with global and gene-specific methylation variations.

In tumors, a global decrease in DNA methylation (hypomethylation) of the genome is accompanied by a region- and gene-specific increase of methylation (hypermethylation) in the normally unmethylated promoter-associated CpG islands, which is often associated with transcriptional silencing of the associated genes. While the contribution of genetic factors to carcinogenesis such as the high-penetrance germ line mutations has long been recognized, it has become evident that epigenetic changes leading to transcriptional silencing of tumor suppressor genes constitute an at least equally contributing mechanism. DNA methylation can act as one of the ‘two hits’ necessary to inactivate the two alleles of a gatekeeper tumor suppressor gene to enable oncogenic progression and can have the same functional effect as a genetic mutation or deletion. Genes involved in resisting cell death, sustaining proliferative signaling, evading growth suppressors, invasion and metastasis enabling replicative immortality, and inducing angiogenesis represent the hallmarks of cancer and have been found to be inappropriately inactivated by DNA methylation.

Aberrant DNA methylation patterns are therefore probably not only a consequence or by-product of malignancy, but can contribute directly to the cellular transformation. It has been extrapolated that aberrant promoter methylation is initiated at ~1 % of all CpG islands and as much as 10 % become methylated during the multistep process of tumorigenesis.

Whole-methylome analysis has become possible due to the huge progress that has been made in the development of array- or next generation sequencing (NGS) based technologies. The Infinium® Human Methylation 450K BeadChip (Illumina Inc., CA, USA), allows the simultaneous quantitative monitoring of the methylation status of more than 450,000 CpG positions and proposes currently the best compromise between throughput, cost, analysis time, resolution and accuracy for genome-wide DNA methylation analysis. However, preprocessing of these data requires many filtering, correction and normalization steps and no consensus and evaluated approach exists to perform all these operations. Furthermore, the Infinium BeadArray uses two different assay chemistries, called Infinium I and Infinium II, whereby Infinium I uses two bead types (for methylated and unmethylated DNA), and Infinium II employs a single bead type, to detect CpG methylation. Several studies reported a shift in the distribution of methylation values measured with each kind of probes [5-6]. This shift may cause a bias in the analysis if both kinds of signals are merged as a unique source of methylation measurement.

In this study we propose a complete preprocessing pipeline for Illumina Infinium 450K data and compare it to alternative approaches. We also propose an original and robust quantile based sample normalization approach that performs both robust sample normalization and efficient Infinium I/Infinium II bias correction. The results of this study have previously been published by Touleimat and Tost in [4].

2 Materials and methods

We used Infinium methylation 450K data from three different studies to analyze and compare the effects of our preprocessing pipeline. These data correspond to tumoral vs. non tumoral studies in three different solid cancers. To validate the methylation values obtained after the application of various variants of our pipeline, we selected from one dataset a subset of differentially methylated probes and used the highly quantitative pyrosequencing technology to obtain a precise value of their “true” methylation status. We also compared the preprocessed signals of various biological categories of Infinium I and Infinium II probes.

The whole preprocessing pipeline is implemented as R scripts that are freely available upon request.

3 Results

We highlight in the three different cancer related datasets the shift in the distribution of Infinium I / Infinium II methylation signals. Direct analysis of these two kinds of signals would correspond to the comparison of probes from two different technologies with different signal distributions. We also show that Infinium I signals are less subjected to experimental variations between samples compared to Infinium II signals. Infinium I signals also cover a wider quantitative range of methylation values than Infinium II signals. Based on these observations we developed an original and robust quantile normalization approach for sample normalization adapted from a subset quantile normalization (SQN) procedure developed in the context of array-based gene expression analysis, where gene-expression signal is normalized with the help of control probes. In our context the reference quantiles are computed on the basis of Infinium I signals only and are further used to estimate Infinium II reference quantiles. Due to the large disequilibrium between Infinium I and Infinium II probe numbers, the Infinium I reference quantiles are “extended” to match Infinium II probes numbers. We also take into account the large imbalance in the proportions of Infinium I and Infinium II probes covering the different CpG and gene-sequence regions by performing the SQN approach by “probe annotation categories”. Our robust SQN approach performs at the same time an efficient between sample normalization and corrects the Infinium I/II signal bias.

In order to obtain the most accurate and robust DNA methylation data, we implemented a complete preprocessing pipeline for the Illumina 450K methylation array that answer, to our knowledge, all the current issues associated to this array. This pipeline is composed of four main steps:

1. Quality Control: batch effect detection (PCA), sample and signal quality estimation and filtering.

2. Probe filtering: filtering of probes with signal variation related to highly variable SNP and/or the filtering of probes associated to allosomal chromosomes.

3. Signal correction: color bias and background subtraction (*lumi* package in *Bioconductor*).

4. Subset based Quantile Normalization: Infinium I/ Infinium II shift correction and between sample normalization.

Steps 1-2 have been developed by using generic functions in R and step 3 is based on functions provided by the *lumi* package in Bioconductor. However, these steps have never been associated as a unique and complete preprocessing pipeline. The last step of the pipeline, the SQN approach, based on a functional annotation system is an original approach, which has, to our knowledge, not been used in the context of DNA methylation data analysis.

In order to evaluate our SQN approach, we also performed alternative approaches for the fourth step of the pipeline: (1) no normalization, (2) classical quantile normalization, (3) peak-based correction [3] followed by quantile normalization, (4) subset quantile normalization with a unique set of reference quantiles computed from Infinium I signals, (5) and finally the SQN approach we developed, where probes, grouped and normalized by “functional categories”. Two variants of this last approach where tested according to the way of defining probe categories: by using either (5a) CpG coverage related annotations (non covered, Island, N_Shore, N_Shelf and S_Shelf) or (5b) by using gene sequence coverage related annotations (non covered, TSS200, TSS1500, 5’UTR, 1st exon, 3’UTR and gene body), both provided by Illumina (CA, USA).

We compared the Infinium 450K data processed with the different normalization variants to the pyrosequencing data, for selected differentially methylated probes. These comparisons, showed that our normalization approach provided the closest methylation intensities to the pyrosequencing based values, which could be considered as the “gold standard” for DNA methylation analysis. The comparison of methylated signals between Infinium I and Infinium II probes associated to the same biological categories proved that our normalization approach provided the most comparable category related methylation variations.

4 Conclusion

Our work shows how a complete and dedicated preprocessing pipeline can solve and correct the current issues raised by Methylation Illumina 450K array users. Furthermore, this pipeline scale well to the high density Illumina methylation 450K array and achieve the whole process in a reasonable time (less than 30 minutes on a classical desktop computer for around 40 samples). Thus, with this robust and relevant preprocessing pipeline, one will only put effort in the biological analysis of the data for methylation signature identification or whole methylome dynamic analysis for example.

The 450K Infinium BeadArray has been released in 2011 and has already revolutionized the field of DNA methylation analysis as it provides for the first time to analyze a reasonable number of CpGs in the human genome at a reasonable price. The highly standardized and automated protocol further permits for the first time the DNA methylation analysis in large cohorts. Although genome-wide sequencing protocols are nowadays available they require important financial resources. Further bioinformatic analyses of these genome-wide DNA methylation sequencing data is currently a major challenge and requires dedicated bioinformatic resources which are not available in most places putting the 450K array in good place for future DNA mwthylation analysis.

Acknowledgements

The research leading to these results has received funding from the EU's Seventh Framework Program (FP7/2007-2013) under grant agreement FP7-HEALTH-2009-241669 (CAGEKID) and from Gen2Phen European program. The authors thank F. Busato and N. Mazaleyrat for performing the Pyrosequencing assays, C. Besse and D. Delafoy for performing the DNA methylation arrays and A. Lemaçon for the integration of the SNP information from the HapMap project in our pipeline. The authors would also like to thank L. Liang from the Harvard School of Public Health (MA, USA) for his valuable advice and comments in evaluating and improving the pipeline.

References

[1] Tost J. DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. Mol. Biotechnol. 44(1), 71–81 (2010).

[2] How Kit A., Myrtue Nielsena H., Tost J. DNA methylation based biomarkers: Practical considerations and applications. *Biochimie*, 94(11): 2314–2337, 2012.

[3] Deng D, Liu Z, Du Y. Epigenetic alterations as cancer diagnostic, prognostic, and predictive biomarkers. *Adv. Genet*. 71, 125–176 (2010).

[4] N. Touleimat and J. Tost. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, 4(3):325-341, 2012.

[5] M. Bibikova, B. Barnes, C. Tsan et al. High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288-295, 2011.

[6] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, F. Fuks. Evaluation of the infinium methylation 450K technology. *Epigenomics*, 3(6): 771–784, 2011.

[7] D. Wang, L. Yan, Q. Hu et al. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* 28(5):729–730, 2012.

Draft genome of the first termite *Zootermopsis nevadensis*

Comparative analysis of a new eusocial and hemimetabolous insect

Nicolas TERRAPON[1], Cai LI[2], Ann Kathrin HUYLMANS[1], Erich BORNBERG-BAUER[1], Jüdith KORB[3] and Jürgen LIEBIG[4]

[1] Institute for Evolution and Biodiversity, Westfälische Wilhelms-Universität, Hüfferstr. 1, D48149 Münster, Germany
n.terrapon@wwu.de, ina-h@wwu.de, ebb@wwu.de

[2] Beijing Genomics Institute Shenzen, 518083, China
licai@genomics.org.cn

[3] Behavioral Biology, University of Osnabrück, Barbarastr. 11, D49076 Osnabrück Germany
Judith.Korb@biologie.uni-osnabrueck.de

[4] School of Life Sciences, Arizona State University, P.O. Box 874501, AZ 85287 Tempe, USA
Juergen.Liebig@asu.edu

Abstract *Termites have substantial economic and ecological impact worldwide. They are also the oldest organisms living in complex societies, having evolved a caste system independent of that of eusocial Hymenoptera (ants, bees, wasps). Here we provide the first genome sequence for a termite, Zootermopsis nevadensis, which represents the phylogenetically deepest rooted insect group (Blattodea) to date. As a consequence, this genome offers a new insights on key protein families for insect evolutionary success. Further, several protein families over-expressed in male reproductives have also shown significant sign of expansion in the termite lineage, which suggest an important evolution in mating biology.*

Keywords Genome sequence, Phylogeny, Insect-specific families, Comparative genomics

Esquisse du génome d'un premier termite *Zootermopsis nevadensis*

Analyse comparative pour un nouvel insecte social et hémimétabole

Résumé *Les termites ont un impact considérable sur l'économie et l'écologie de notre planète. Ils sont également les plus anciens organismes à vivre en sociétés complexes, ayant évolué vers un système de castes indépendant de celui des hyménoptères (fourmis, abeilles, guêpes). Nous présentons ici le premier génome d'un termite, Zootermopsis nevadensis, qui représente le groupe (Blattodea) le plus profondément enraciné à ce jour dans la phylogénie des insectes. En tant que tel, ce génome offre une perspective nouvelle sur les familles protéiques ayant joué un rôle-clef dans le succès évolutif des insectes. De plus, plusieurs familles protéiques surexprimées chez les mâles reproducteurs montrent également des signes significatifs d'expansion chez le termite, ce qui suggère une importante évolution dans biologie de la reproduction.*

Mots-clés Génome séquencé, Phylogénie, Familles spécifiques des insectes, Génomique comparative

1 Introduction

Les termites sont considérés comme le second groupe d'insectes le plus important du point de vue écologique après les abeilles. Grâce à leur diète (xylophagie), les termites participent à l'entretien et l'enrichissement des sols dans les prairies et les forêts des zones tropicales où ils représentent plus de 10% de la biomasse [1]. D'un autre côté, les termites produisent de grandes quantités de gaz à effet de serre pour une émission estimée à 2-5% de la production de méthane annuelle. L'impact économique des termites est également considérable puisque environ 10% des ∼3000 espèces décrites sont considérées comme nuisibles car elles détruisent les récoltes, les pâturages et les immeubles pour un coût moyen de 20 milliards de dollars par an à travers le monde.

Le tout premier génome de termite a été séquencé. Ce génome offre une nouvelle perspective sur l'évolution des insectes étant donnés sa position phylogénétique et les traits qu'il partage avec les insectes déjà séquencés.

Les termites sont des formes spécialisées de cafards, la lignée d'insectes la plus profondément enracinée par rapport aux génomes d'insectes actuellement disponibles [1]. Les termites sont également les plus anciens insectes complètement sociaux, vivant en colonies de 100 à 1 000 000 individus des deux sexes, différenciés en castes (Fig. 1) :

- les ouvriers (seule caste composée d'individus immatures – larves et nymphes – pouvant encore évoluer vers les suivantes) construisent et nettoient la termitière, s'occupent des oeufs et nourissent la colonie ;
- les soldats sont chargés de protéger la colonie à l'aide d'attibuts morphologiques ou chimiques ;
- les *ailés* sont voués à quitter la termitière pour fonder une nouvelle colonie ;
- le couple royal se consacre entièrement à la reproduction (fécondations répétées – important au regard des résultats présentés en section 6 – dans le copularium où ils sont confinés). La reine dont l'abdomen dépasse parfois les 10 cm, pondra plusieurs millions d'oeufs au cours de sa vie qui peut durer plus d'une vingtaine d'années.

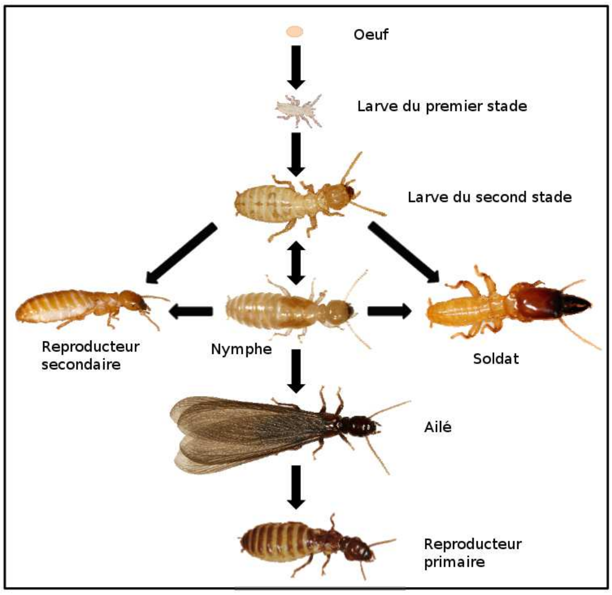


Figure 1. Développement/castes du termite *Z. nevadensis*.

La socialité a évolué indépendamment chez les termites et les hyménoptères et il est maintenant possible de comparer l'évolution vers des sociétés avec une détermination sexuelle diplo-diploïde et haplo-diploïde. De plus, le termite est le troisième génome d'insectes hémimétaboles, c'est-à-dire avec un cycle de vie caractérisé par une métamorphose incomplète. Ce groupe n'a pas reçu jusque-là autant d'attention que les holométaboles (diptères, coléoptères, hyménoptères, etc.) et pourtant chaque nouveau génome d'hémimétabole apporte un regard nouveau sur l'évolution de la métamorphose.

Le génome du termite du bois humide *Zootermopsis nevadensis* sera bientôt disponible. Nous présentons brièvement ici les données génomiques et transcriptomes (section 2), l'annotation fonctionnelle des protéines et les espèces d'insectes utilisées comme références pour les différentes analyses (section 3), l'analyse phylogénétique (section 4), quelques familles protéiques d'intérêt (specifiques des insectes – section 5) et une analyse de génomique comparative menée pour identifier les caractéristiques évolutives spécifiques au termite comparé aux arthropodes de référence (section 6).

2 Génome et transcriptomes

2.1 Séquençage/assemblage du génome et prediction de protéines

Les individus de *Z. nevadensis*, prélevés sur une côte Californienne, ont été traités par séquençage à très haut débit (Illumina Hiseq 2000) au Beijing Genomics Institute (BGI). Les libraries de fragments produites ont été assemblées avec le logiciel SOAPdenovo en 493 Mbp pour une taille de génome estimée à 562 Mbp (soit un génome à 88% complet).

La prédiction de 17 737 gènes/protéines s'est appuyée sur une combinaison de :

– données transcriptomiques : 13 individus différents (section 2.2) ont été séquencées par des techniques à haut débit (RNA-seq), le logiciel Tophat a été utilisé pour l'alignement sur le génome et Cufflinks pour reconstruire les transcrits.
– modèles de gènes : les gènes de l'humain *Homo sapiens*, la mouche *Drosophila melanogaster*, l'abeille *Apis mellifera* et la fourmi *Harpegnathos saltator* ont été alignés sur le génome via TBLASTN et les modèles créés avec GeneWise.
– prédictions *de novo* : en utilisant les logiciels Augustus et SNAP (après avoir masqué les éléments répétés du génome) pour former des modèles retenus en cas de resultats significatifs avec BLASTP contre Swiss-prot.

2.2 Transcriptomes RNA-seq de 13 individus

Pour ce projet, les transcriptomes de 13 individus de castes et sexes différents ont été sequencés. Les échantillons ont été collectés pour des oeufs et des ouvriers (larves et nymphes) sans différenciation de sexe, tandis que les mâles et femelles ont donné lieu à des échantillons distincts pour les soldats, les ailés et les reproducteurs primaires et secondaires. Les reproducteurs primaires sont les fondateurs d'une colonie et sont donc rares. Les reproducteurs secondaires sont des nymphes qui sont appelées à régner et se développent en reine ou en roi suite à la mort de leur prédécesseur. Chez les termites, les reproducteurs primaires et secondaires peuvent être sexuellement actifs (ovaires produisant des oeufs ou testicules produisant du sperme) ou inactifs (aucun signe d'oocytes ou petits testicules). Les transcriptomes ont été reconstruits pour des reproducteurs primaires inactifs et des reproducteurs secondaires actifs et inactifs.

Après séquençage, les fragments alignés ont permis le calcul de valeurs d'expression pour chaque protéine sous la forme de RPKM (*Reads Per Kilobase per Million mapped reads*). Ces valeurs ont été normalisées à travers les échatillons en suivant la méthode *Trimmed Mean of M-values* (TMM). Nous avons alors appliqué un clustering par *K-means* pour identifier les protéines ayant des profils d'expression similaires dans les différents individus et détecter les gènes différentiellement exprimés.

3 Espèces de référence et annotation fonctionnelle

3.1 Sélection d'espèces de référence

Six organismes de référence ont été utilisés pour des études préliminaires en phylogénie et en génomique comparative : *Tribolium castaneum* (*red flour beetle*) [2], deux insectes sociaux – l'abeille *Apis mellifera* [3] et la fourmi *Harpegnathos saltator*[4]–, les deux espèces hémimétaboles actuellement disponibles – le puceron *Acyrthosiphon pisum* [5] et le pou *Pediculus humanus* [6]–, et enfin pour un point de vue extérieur aux hexapodes/insectes, la daphnie (crustacé) *Daphnia pulex* [7]. La phylogénie de ces espèces (Fig. 2) intègre également la position vraisemblable du termite (cf. section 4).

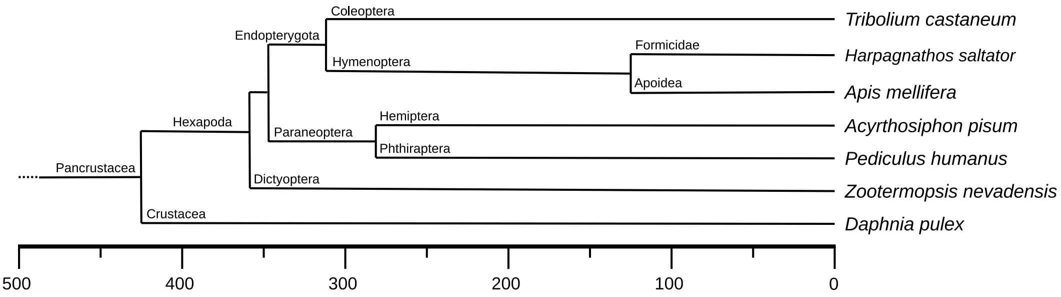


Figure 2. Phylogénie des espèces de référence et du termite *Z. nevadensis*. Les dates estimées de spéciation sont extraites de la publication du génome de l'abeille [3] et de la litterature pour le termite [8].

3.2 Annotation des protéines

L'annotation fonctionnelle des protéines a été réalisée dans les espèces de référence ainsi que chez le termite. Les domaines protéiques, unités structurales, fonctionelles et évolutives des protéines, ont été identifiés via la base de données Pfam et la métabase Interpro. À partir des domaines, des annotations de la *Gene Ontology* (GO) ont pu être déduites. Une source additionnelle d'annotation automatique concerne les voies métaboliques de la base KEGG. Toutes ces annotations ont été utilisées dans l'analyse de génomique comparative pour tenter d'identifier les caractéristiques spécifiques aux termites, ou communes aux insectes sociaux ou hémimétaboles. Nous nous concentrons ici sur les annotations en domaines Pfam, notamment utilisées pour créer des *familles protéiques* en regroupant toutes les protéines exhibant une composition identique en domaines (sans prendre en compte ni les répétitions ni l'ordre). Toutefois, une vue plus détaillée des groupes d'orthologues, *sous-familles*, a été générée par l'algorithme orthoMCL et l'intégration plus récente du termite dans la base de données d'orthologie orthoDB [9].

3.3 Qualité du génome

Un premier contrôle pour la qualité du génome avait pour objectif la détection chez le termite d'un quelconque biais qui aurait pu remettre en question l'assemblage ou la prédiction des protéines. Dans ce but, nous avons notamment comparé l'annotation en domaines Pfam (Table 1), mais également d'autres indices tels que le taux de fragmentation des domaines et d'autres annotations fonctionnelles (données non présentées, ex. KEGG), la longueur moyenne des protéines et la proportion de protéines *clusterisées* par orthoMCL (Fig. 3). On constate que le termite se place parfaitement dans la moyenne des autres génomes d'insectes de référence (dont la publication est plus ancienne) ce qui, sans être une preuve irréfutable, est un bon indicateur de la qualité du génome.

Espèce	Protéines	Occ. dom.	Fam. Pfam	Cov. prot.	Impl. AA	Multidom.
T. castaneum	16631	19015	3395	10397	34,0%	1697
H. saltator	17191	15700	3409	8927	31,2%	1476
A. mellifera	10660	15747	3432	8288	35,3%	1602
P. humanus	10769	14654	3310	7727	33,8%	1486
A. pisum	33267	20809	3308	13028	24,8%	1467
Z. nevadensis	17737	17505	3460	9201	31,8%	1578
D. pulex	30899	21759	3462	13651	28,0%	1553

Table 1. Statistiques de l'annotation en domaines Pfam chez les espèces de référence et le termite. Pour chaque espèce, "Protéines" indique le nombre total de protéines, "Occ. dom." le nombre total d'occurrences, "Fam. Pfam" le nombre de familles Pfam uniques, "Cov. prot." le nombre de protéines avec au moins un domaine, "Impl. AA" le pourcentage moyen de la protéine couvert par des domaines, et "Multidom." la quantité de compositions distinctes en domaines.

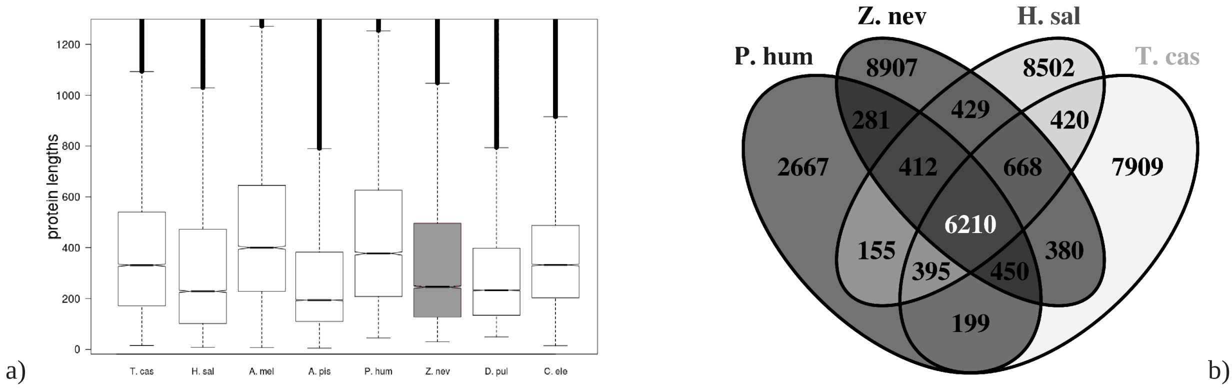


Figure 3. a) Longeur moyenne des protéines dans les espèces de référence et le termite. b) Diagramme de Venn des protéines orthologues d'après orthoMCL pour le pou, le termite, la fourmi et *Tribolium*.

4 Analyse phylogénétique

La phylogénie relative des groupes profondément enracinés dans l'arbre des insectes est toujours sujette à controverse. Outre l'échantilonnage taxonomique réduit, la principale cause est vraisemblablement une ancienne et rapide radiation de ces groupes. Nous avons donc mené des investigations pour déterminer la position du termite dans l'arbre phylogénétique des insectes (Fig. 2). Nous avons tout d'abord retenu l'ensemble des sous-familles d'orthologues 1:1 (une copie par espèce), créé les alignements multiples de ces familles avec MAFFT et nettoyé les alignements avec Gblocks. Ces alignements ont ensuite été concaténées en un super-alignement d'acides aminés duquel nous avons déduit un super-alignement de nucléotides (deux premières positions uniquement). Ces deux super-alignements ont été utilisés pour calculer des topologies à l'aide de deux approches : *maximum likelihood* avec morePhyML [10] et bayésienne avec Phylobayes [11]. Les résultats obtenus (acides aminés/nucleotides, ML/bayésien) présentent quatre topologies distinctes entre le termite, le pou et le puceron, aucune correspondant à l'arbre attendu (Fig. 4). L'un des biais vient vraisemblablement du puceron dont l'évolution a été rapide, avec de nombreux paralogues mal classifiés (observation personnelle), et les modèles de gènes souvent incomplets.

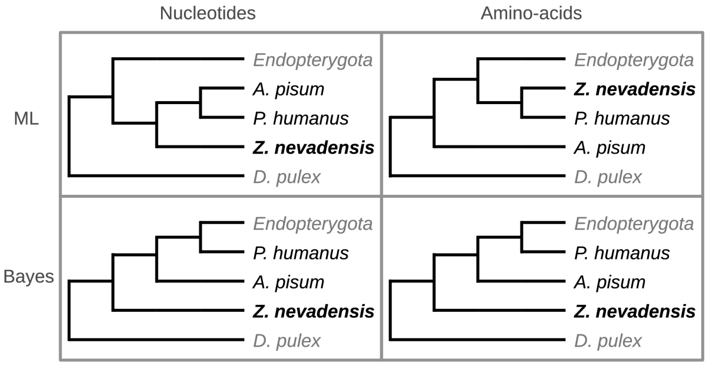


Figure 4. Incohérence des topologies préliminaires pour différents codages et méthodes phylogénétiques.

Plusieurs solutions ont été testées pour contourner ces limitations, par exemple en ne retenant que les familles où tous les membres ont une composition identique en domaines (pour éviter les modèles de gène incomplets) ou les familles avec l'évolution la plus lente, mais sans succès. Le principal problème semble être la courte période de temps entre les divergences de ces groupes en comparaison à la grande distance phylogénétique à l'outgroup *D. pulex*. Par conséquent, nous avons remplacé *D. pulex* par les trois groupes (Thysanura, Archeognatha et Diplura), avec un enracinnement plus profond que le termite mais appartenant aux *Hexapodes* (Fig. 5). Ces groupes ont été selectionnés pour leur grand nombre de protéines actuellement disponible dans la base de données du NCBI, malgré l'absence de génomes complets. Nous avons identifié par Blast réciproques les orthologues pour 16 des familles protéiques initiales et reconduit nos analyses phylogénétiques. Les quatre topologies obtenues par ce jeu de données sont toutes en accord entre elles et avec la position attendue du termite dans la phylogénie des insectes.

5 Larges familles de gènes spécifiques au insectes

Étant donné la position privilégiée du termite au sein de la phylogénie des insectes, nous avons porté une attention particulière aux familles de gènes ayant joué un rôle critique dans le succès écologique de ces espèces. Nous nous sommes intéressés en particulier aux larges familles de gènes spécifiques des insectes, c'est-à-dire totalement absents des génomes d'arthropodes ayant divergé avant l'insecte ancestral (ex. *D. pulex*). Deux familles répondent à ces critères et présentent également un fort intérêt fonctionnel, lié aux premiers stades du développement par exemple pour les gènes Osiris.

5.1 les gènes *Yellow*

Ils sont caractérisés par la présence d'un domaine MRJP (pour *Major Royal Jelly Protein*) dont la fonction n'est pas clairement comprise [12]. On distingue environ 12 sous-familles de gènes *Yellow*, dont l'une, nommée

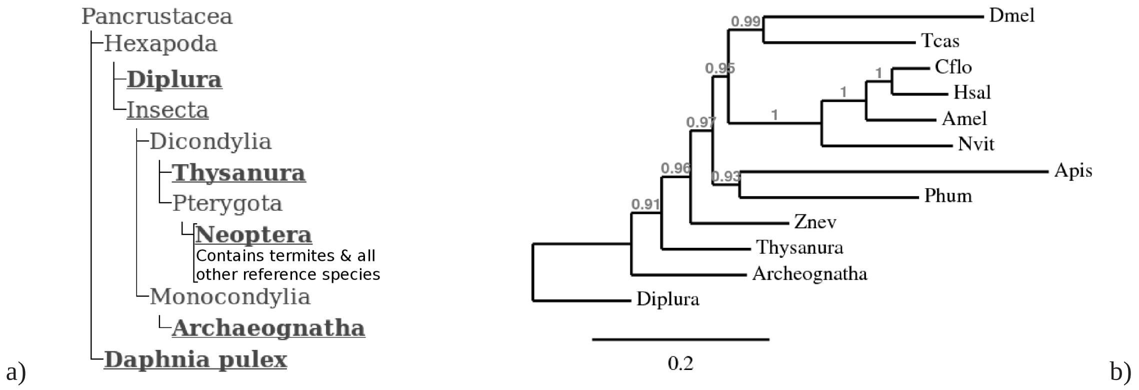


Figure 5. a) Classification taxonomique du NCBI de la distante daphnie et des nouveaux outgroups. b) Phylogénie finale obtenue sur les séquences d'acides aminés par morePhyML.

MRJP comme le domaine, a largement été dupliquée chez les hyménoptères et joue un rôle important dans les comportements sociaux de l'abeille notamment. Parmi ces sous-familles, certaines forment une zone de micro-synténie conservée chez tous les insectes, avec au moins 6 protéines d'après les hypothèses les plus récentes [12]. Le termite confirme l'importance de la conservation de cette zone qui contiendrait vraisemblablement 7 protéines dans l'ancêtre commun. À l'exception du pou qui a perdu 3 des copies synténiques, la plupart des espèces n'en ont perdu qu'une, comme c'est le cas chez le termite.

5.2 les gènes *Osiris*

Initialement décrit chez *D. melanogaster* avec 24 protéines, 21 présentent une très forte micro-synténie à travers la plupart des génomes d'insectes, y compris les génomes hémimétabolés. La plus faible conservation est observée chez le pou avec seulement 13 gènes dont 8 synténiques. Nous avons observé chez *Z. nevadensis*, la présence de 17 gènes Osiris, dont 11 sont synthéniques, suggérant la création/l'acquisition du gène et sa fréquente duplication avant l'ancêtre commun au termite et aux espèces de référence. De plus, 12 des protéines du termite présentent une forte sur-expression dans les transcriptomes d'oeufs, très similaires à ce qui a pu être observé chez *D. melanogaster* (depuis les embryons de 12-18h aux larves de 2 jours d'après Flybase). Malgré le fait que l'on ne sache que très peu de choses sur la fonction des gènes Osiris, leur expression dans un grand nombre de tissus à l'exception du système nerveux semble indiquer un rôle important pour le développement. Le termite offre alors un complément d'information important et des corrections aux hypothèses formulées sur la spécificité phylogénétique de certaines copies/sous-familles [13] (Fig. 6).

6 Expansion de familles protéiques

Nos investigations sur l'évolution du répertoire protéique ont notamment révélé l'expansion de plusieurs familles de protéines chez le termite. En utilisant la classification des protéines en familles (basée sur la composition en domaines – cf. section 3), nous avons comparé la taille des familles chez le termite à chacune des espèces de référence indépendamment, à l'aide de tests exacts de Fisher. Des p-valeurs significatives, unidirectionnelles (expansion ou contraction) et obtenues par rapport à une majorité d'espèces de référence, ont permis d'identifier les familles suivantes :

- BTB-BACK-KELCH (tridomaine) et KELCH (monodomaine – vraisemblablement fragments de protéines tridomaines car une majorité sont situés aux extrémités de scaffolds). Nous avons comptabilisé respectivement 37 protéines tridomaines et 20 monodomaines chez le termite contre seulement 4 à 10 et 0 à 10 dans les espèces de référence, à l'exception du puceron qui a subi une expansion encore plus forte avec 78 et 65 protéines respectivement.
- PKD_channel protéines (monodomaines), avec 10 protéines chez le termite et 1 à 3 dans les espèces référence à l'exception de *D. melanogaster* avec 6 protéines.
- *Seven-in-absentia* (Sina) protéines (monodomaines) avec 33 protéines chez *Z. nevadensis* mais seulement 1 à 4 dans les espèces de référence, excepté *T. castaneum* avec 16 protéines.

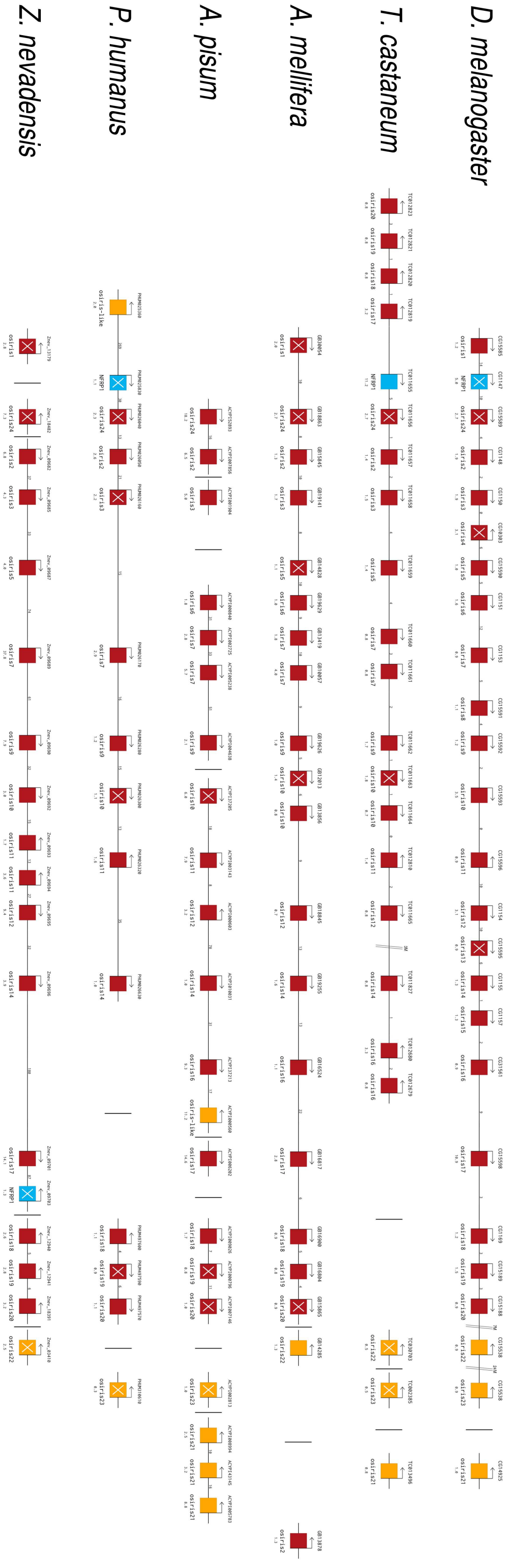


Figure 6. Gènes Osiris chez le termite et 5 des espèces de référence. Les gènes répartis sur différents scaffolds sont représenté dans l’ordre observé chez *D. melanogaster*. Les carrés représentent les gènes (rouge pour les Osiris a priori synthéniques, jaune pour les autres, bleu pour le gène NFRP *niché* dans la zone synthénique), une flèche pour la direction de transcription, un trait vertical indique les différents scaffolds.

En croisant ces informations avec les données d'expression des transcriptomes de *Z. nevadensis*, nous avons observé une sur-expression significative (tests exacts de Fisher) de ces familles chez les reproducteurs mâles. Cela concerne :

– 26 des 37 BTB-BACK-KELCH et 16 des 20 protéines monodomaines KELCH. De plus, l'analyse d'expression différentielle a révélé la sur-expression chez les mâles reproducteurs des 4 protéines BACK-KELCH, de 5 des 6 protéines BTB-KELCH et de 3 protéines BTB-BACK (sur 9 exprimées, 12 au total). Ces protéines semblent être dues à des modèles de gènes incomplets de protéines tridomaine, ce qui renforcerait encore l'expansion et l'expression spécifique de cette famille ;
– 8 des 10 protéines PKD_channel ;
– 7 des 23 protéines Sina exprimées (sur 33) tandis que 11 sont sur-exprimées chez les femelles reproductrices ;

Nous avons également pu confirmer par des analyses phylogénétiques que, pour chacune des familles, les protéines différentiellement exprimées apparaissent dans un unique groupe monophylétique de paralogues (issus d'une protéine ancestrale orthologue à celle des espèces de référence). Ces observations, associées aux annotations fonctionnelles issues des espèces de référence et des domaines, suggèrent une co-expansion au cours de l'évolution de *Z. nevadensis* pour le développement de nouveaux mécanismes liés à la reproduction chez les termites, principalement à la spermatogenèse. Cette découverte est particulièrement pertinente étant donné la nécessité de reproduire du sperme régulièrement pour le roi lors des reproductions (section 1), répétées au cours de la vie du couple royale mais espacées par de longues périodes.

Références

[1] D.A. Grimaldi and M.S. Engel, *Evolution of the Insects.*

[2] S. Richards, *et al.*, The genome of the model beetle and pest Tribolium castaneum. *Nature*, 452 :949-955, 2008.

[3] G.M. Weinstock, *et al.*, Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, 443(7114) :931-949, 2006.

[4] R. Bonasio, *et al.*, Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. *Science*, 329 :1068-1071, 2010.

[5] "The International Aphid Genomics Consortium", Genome sequence of the pea aphid Acyrthosiphon pisum. *Plos Biol.*, 8(2) :e1000313, 2010.

[6] E.F. Kirkness, *et al.*, Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *PNAS*, 107(27) :12168-12173, 2010.

[7] J.K. Colbourne, *et al.*, The ecoresponsive genome of Daphnia pulex. *Science*, 331(6017) :555-561, 2011.

[8] H.O. Letsch, *et al.*, Insect phylogenomics : results, problems and the impact of matrix composition. *Proc. R. Soc. B-Biol. Sci* 279 :3282-3290, 2012.

[9] R.M. Waterhouse, E.M. Zdobnov, F. Tegenfeldt, J. Li and E.V. Kriventseva, OrthoDB : the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic acids research*, 39(Database issue) :D283–D288, 2011.

[10] A. Criscuolo, morePhyML : improving the phylogenetic tree space exploration with PhyML 3. *Mol Phylogenet Evol.*, 61(3) :944–948, 2011.

[11] N, Lartillot, T. Lepage, S. Blanquart, PhyloBayes 3 : a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17) :2286–2288, 2009.

[12] L.C., Ferguson, *et al.*, Evolution of the insect yellow gene family. *Molecular biology and evolution*, 28, 257–272, 2011.

[13] Shah, N., *et al.*, Evolution of a Large, Conserved, and Syntenic Gene Family in Insects, *G3 : Genes— Genomes— Genetics*, 2, 313–319, 2012.

Unraveling evolutionary mechanisms driving metabolic diversity

Pablo CARBONELL[1,2] and Pierre PARUTTO[1,2]

[1] Univ. Evry, iSSB, F-91000 Évry, France
[2] CNRS, iSSB, F-91000 Évry, France
pablo.carbonell@issb.genopole.fr

Abstract *How enzymes have acquired their specificity during the course of evolution is a process still not fully understood. In extant metabolic networks some enzymes are highly specialized, while many others are reported with the ability to catalyze several distinct reactions, a property known as enzyme promiscuity. Understanding such metabolic adaptation processes will assist us into the conception of advanced enzymatic capabilities for next-generation synthetic biology devices. To that end, we model here two possible mechanisms driving enzyme evolution, the progressive and the regressive models, correlating the results of their simulations to the promiscuity distribution as it is observed through phylogenetic studies of modern metabolic networks.*

Keywords Enzme, promiscuity, evolution.

1 Introduction

Understanding which mechanisms facilitate the acquisition of new enzyme functions is a problem that is central to bioengineering. By using directed evolution, protein engineers are able to evolve new functions either from parent enzymes or from *de novo* scaffolds. These new functions are of special interest in synthetic biology since they can provide novel ways to engineer organisms so that their biosynthetic or biodegradation routes can process chemical products with potential industrial, health or environmental applications. Similarly, new enzymes are needed to be evolved as modular parts of next-generation synthetic biological circuits such as for implementing a biochemical transformation of some chemical of interest into an inducer regulating gene expression for sensing biomarkers and in metabolic regulation.

However, a major hindrance in the study of enzyme specificity is the lack of a rigorous definition of enzyme promiscuity. We have recently proposed a new way to measure enzyme promiscuity based on the reaction signature diversity [1,2]. In contrast with other definitions based on the classification of enzymatic reactions through the EC number coding, which is prone to ambiguity, in the reaction signature diversity definition we measure how close two enzymatic reactions are by mapping the chemical space into the reaction space, a metric space with a well-defined algebra. This method thus allows us to mine efficiently the biochemical data available in metabolic databases in order to fully characterize metabolic networks from the promiscuity of their enzymes.

Notably, we have shown that enzyme promiscuity plays a prominent role in the study of the evolution of metabolic networks [3]. A distinctive property of our given definition of promiscuity, which is based on the chemical diversity of catalytic reactions that enzymes can process, is that it does not depend on protein sequence comparisons. Although many enzyme families and superfamilies share structural and functional features, distant evolutionary relationships might be difficult to identify solely from global sequence comparison. Our method, therefore, should be regarded as complementary to sequence-based methods. In that manner, the same evolutionary trend in enzyme promiscuity was observed in the tree of life independently of the way pairwise distances between species were measured, either based on sequence data or on pure taxonomic hierarchy [3]. We found, however, that our chemical approach can be potentially closer to physiology and thus to explain the organism response to stress due to some environmental pressure. For instance, we observed higher chemical diversity in extremophiles, a group of species that can survive under extreme conditions and therefore need a highly adaptable metabolic system.

Here, this definition of enzyme promiscuity is used in order to build and simulate two models embodying the two main proposed mechanisms of enzyme evolution: the retrograde and patchwork models [4]. We have

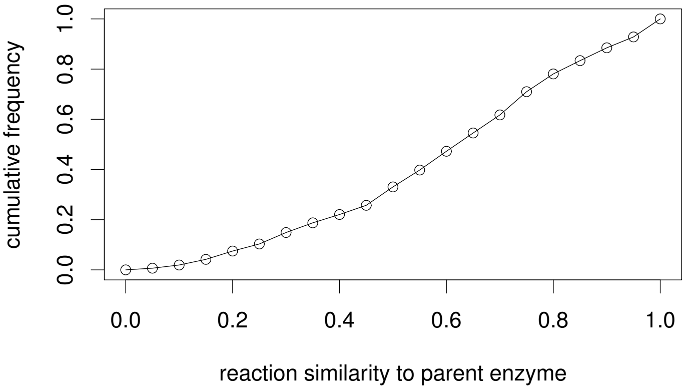


Figure 1. Cumulative probability distribution of a reaction to be present in a promiscuous enzyme.

previously presented a method to simulate how reaction networks are formed and evolved based on the Gillespie algorithm for stochastic simulation [5]. This method can be adapted in order to simulate the evolution of metabolic networks. In our simulations, every time a new evolutionary event arrives, new reactions are added to the network. Whether a new enzyme appears or reactions are added to the current enzymes is given by a probability distribution curve estimated from the observed values of promiscuity in metabolic networks.

2 Models of Enzyme Specifity Evolution

2.1 Reaction Diversity in Extant Enzymes

Metabolic databases such as KEGG or MetaCyc provide valuable information about the probability at large for an enzyme to catalyze two reactions depending on how similar they are. By using our definition of chemical similarity between reactions, we computed similarity between reactions that promiscuous enzymes can catalyze in the KEGG database. Such distribution is given in FIG. 1. As we could expect, two reactions with high similarity are more likely to be catalyzed by the same enzyme. However, cases where reactions with high dissimilarity are present in promiscuous enzymes are also reported, for instance in the case of a multi-domain enzyme. Therefore, such computed distribution provides a general view of the chemical versatility observed in extant enzymes.

In order to perform the simulations in our models, we sampled the distribution shown in FIG. 1 to determine the probability of an event involving two reactions to give birth to a new enzyme through duplication. Such approach, even though it does not consider the mechanisms behind the acquisition of the new enzymatic activity, allows to reproduce at large a similar behavior as the one resulting from natural evolution.

2.2 The Progressive Model

The "progressive model" assumes that originally only few primeval reactions were present in enzymes and then, according to the retrograde model, enzymes slowly acquired during evolution new functions as long as precursors of the new reactions became available in the metabolic network. Based on this model, what we observe in modern enzymes of older origins is a greater reaction repertoire because of the progressive accumulation of newly acquired functions. This model assumes that enzymes acquire promiscuity during the course of evolution as new reactions appear. Such mechanism, thus, should explain the observed fact that older enzymes are on average more promiscuous: they had enough time to accumulate more reactions.

The evolutionary simulation of this model was implemented as a Gillespie algorithm [5]. Following the approach in [6] , we started with a basic set consisting of common currency metabolites and a set of primordial metabolites based on common conserved enzyme fragments, ATP phosphohydrolase was considered the

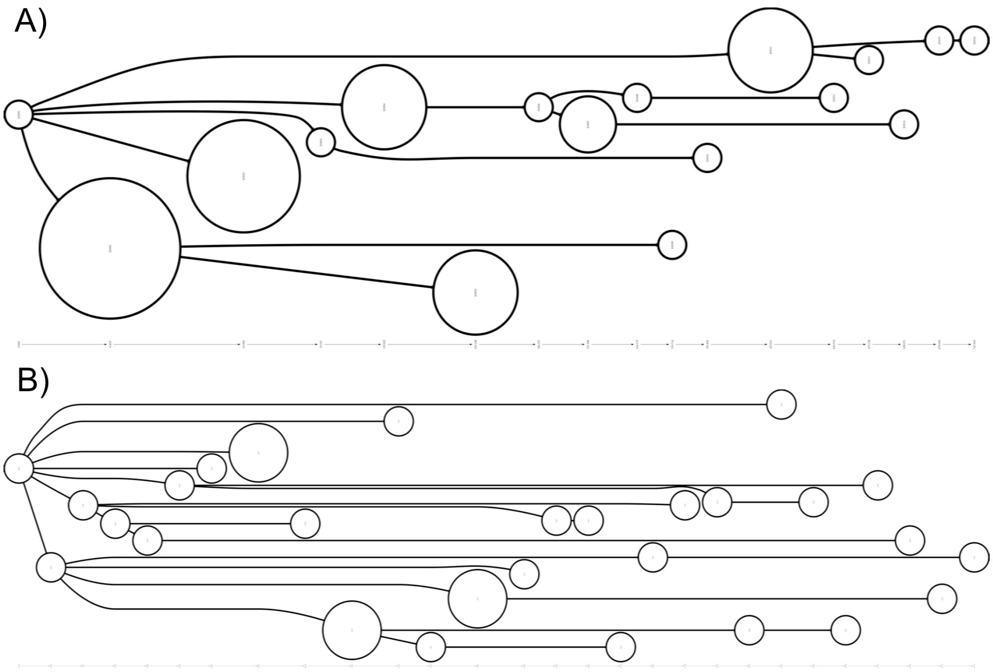


Figure 2. Timeline of enzyme evolution in a simulation of the A) progressive and B) regressive models. Node size represents degree of enzyme promiscuity.

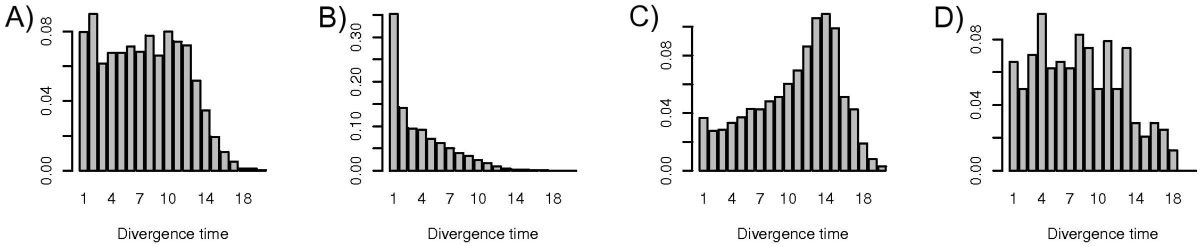


Figure 3. Distribution of divergence times in the simulations of the progressive model for A) non-promiscuous and B) promiscuous enzymes, and the regressive model for C) promiscuous and D) non-promiscuous enzymes.

primeval enzymatic reaction [7] . Following the approach in [6], in order to evolve the network from the list of initial enzymes, we considered at each step of the simulation all reactions in the metabolic network that potentially could proceed based on the current availability of their precursors in the set of current metabolites. Only once all precursors are available in the medium, a new reaction can be formed.

We defined the propensity p_i of a new reaction to be evolved from an enzyme as its chemical similarity to the set of reactions catalyzed by that parent enzyme. Based on that measure, the parent enzyme of a given reaction was determined to be the enzyme with the maximum propensity p_i to evolve that reaction. At each evolutionary event of the simulation:

1. The new reaction to appear is obtained by randomly sampling the reactions with a probability given by $p_i / \sum p_i$.

2. The new reaction is either added to the repertoire of catalytic capabilities of the enzyme as a promiscuous reaction or specializes into a new enzyme evolved from the parent. In order to determine this event, we performed a Monte Carlo sampling of the frequency distribution of chemical similarities in reactions contained in promiscuous enzymes, as it is observed in the KEGG metabolic database, as described in the section below.

3. Time scale can optionally be added to the simulations by considering that evolutionary events occur in time according to an exponential distribution. The next evolutionary event is determined by an exponentially distributed random variable with mean given by $1/\sum p_i$.

2.3 The Regressive Model

The "regressive model" assumes, as in the patchwork model, that ancient enzymes were highly promiscuous and that they became specialized through duplication events during the course of evolution. According to this model, promiscuity would be present in most of the cases in enzymes that had been for longer time present during evolution and therefore are more spread across the tree of life. We start from a set of primeval reactions, that we assume to be already present at some level of latency in the highly promiscuous ancient enzymes.

The propensity p_i of a reaction to become independent of its parent enzyme through specialization into a newborn enzyme is defined as the minimum chemical similarity to the reactions in its parent enzyme. A reaction that it is more distant to the rest of the reactions is more likely to specialize into a new enzyme. At each evolutionary step of the simulation:

1. The new reaction to be specialized was randomly generated based on propensities, as in the progressive model.
2. Reactions in the parent enzyme can either remain in it or become part of the repertoire of the newborn enzyme. Which reactions are contained in the new enzyme and which ones remain in the parent is determined by sampling of the frequency distribution in KEGG, as described in the section below.
3. Time scale can be added as in the progressive model.

3 Results and Conclusions

A typical simulation run is shown in FIG. 2, while a summary for 10^3 simulations based on the KEGG database is shown in FIG. 3. There are basically two distinctive trends: a) promiscuous enzymes in the progressive model are generally of older origin, while b) non-promiscuous (more specific) enzymes appear generally later in time in the regressive model. Therefore, these results show that even though both models have the ability to partially explain the distribution of enzyme functionalities in modern metabolic networks, the way evolution should be traced in the network differs depending on the underlying mechanisms at work behind the generation of enzyme promiscuity. We might infer from these preliminary data that natural pathways emerged as a tradeoff between the evolutionary pressures induced by the constrains associated to each of the two models. Namely, the progressive model has the ability of representing random events associated with the acquisition of new reactions that were selected because of their evolutionary advantage. The regressive model, in turn, provides a closer look into the plasticity and robustness mechanisms of metabolic networks through the emergence of the latent repertoire of promiscuous enzymatic activities.

In our previous studies [3] we showed the relationship between chemical diversity of promiscuous enzymes and their phylogenetic diversity. Here, we have proposed and tested *in silico* two models that can explain such evolutionary traits. These results, thus, provide new clues in order to explain the mechanisms of how enzymatic activities evolve and specialize under environmental pressure. Natural selection is implicitly included in the model, since we generated random events that parallel selective pressure from environmental pressure that resulted in the emergence of a new enzymatic activity. In that way, we obtained a similar trend in our model than in extant networks, even though the mechanisms at play behind were solely given implicitly. Future work, thus, should introduce in our simulations natural events in the history of life, such as the even that gave rise to adaptation into molecular oxygen [8]. Understanding such mechanisms will assist us into the conception of new metabolic capabilities and towards the identification of primeval promiscuous catalytic functions at the core of life's minimal metabolism.

4 Acknowledgements

Funding: Genopole® through an ATIGE grant; ANR Chair de Excellence.

This work was supported by JOBIM 2013.

References

[1] Carbonell, P., Larsson, L. and Faulon, J.L. Stereo signature molecular descriptor. *J Chem Inf Model.*, 53(4):887–897, 2013.

[2] Carbonell, P. and Faulon, J.L. Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics*, 26(16):2012–2019, 2010.

[3] Carbonell, P., Lecointre, G. and Faulon, J.L. Origins of specificity and promiscuity in metabolic networks. *J Biol Chem*, 51: 43994–44004, 2011.

[4] Yamada, T. and Bork, P. Evolution of biomolecular networks — lessons from metabolic and protein interactions. *Nat Rev Mol Cell Bio*, 10(11):791–803, 2009.

[5] Faulon, J.L. and Carbonell, P. Reaction network generation. *Handbook of Chemoinformatics Algorithms*, 317–342, 2010.

[6] Schütte, M., Skupin, A., Segrè, D. and Ebenhöh, O. Modeling the complex dynamics of enzyme-pathway coevolution. *Chaos*, 20(4):045115+, 2010.

[7] Sobolevsky, Y., Frenkel, Z., and Trifonov, E. Combinations of ancestral modules in proteins. *J Mol Evol.* 6: 640–650, 2007.

[8] Raymond, J, and Segrè, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science*, 311(5768):1764–1767, 2006.

Efficient mapping of short peptides on whole proteome database for biomarker discovery

Thomas HUME[1,2], Hayssam SOUEIDAN[3], V. Fabienne WONG JUN TAI[2], Antoine VEKRIS[4], Klaus PETRY[4] and Macha NIKOLSKI[1,2]

[1] Univ. Bordeaux, CNRS / LaBRI, F-33405 Talence, France
{thomas.hume, macha.nikolski}@labri.fr
[2] Univ. Bordeaux, CBiB, F-33000 Bordeaux, France
virginie.wongjuntai@u-bordeaux2.fr
[3] NKI-AVL, Plesmanlaan 121, 1066 CX Amsterdam, Netherlands
h.soueidan@nki.nl
[4] Univ. Bordeaux, INSERM U1049, F-33000 Bordeaux, France
avec@mac.com, klaus.petry@inserm.fr

Abstract Context: In vivo *phage display can be used to simultaneously screen peptide repertoires for peptides presenting specificity for certain interaction and recognition sites, and thus enable biomarker identification. In this work we assume the hypothesis that certain small peptides mimick physiological interactions of proteins that contain a sub-sequence similar to these peptides. A possible solution then involves mapping of the peptide repertoire against the proteome of interest, which is computationnaly challenging. We formulate this problem as simultaneous matching of multiple patterns against multiple strings and propose an algorithmically efficient solution. It is currently accessible at* http://services.cbib.u-bordeaux2.fr/spack/pepteam.php.

Keywords mapping, string matching, peptide, phage display

1 Introduction

The aim of this work is to streamline the biomarker discovery based on the hypothesis of mimed proteins in the context of selection of phage displayed peptides. Phage display is an *in vitro* or *in vivo* technique used to identify relevant protein interaction and recognition sites, first described in 1985 by Smith [1]; it was later shown that antibody fragments could be successfully displayed on phage [2]. The technique is based on the insertion of DNA fragments into bacteriophage genes to create fusion proteins with the foreign sequence in the middle. Viral particles contain the encoding DNA and display the encoded peptide on the surface of their capsizes. Combinatorial libraries of peptides can be produced with complexities of 10^9 and screened to isolate peptides that bind to biological samples ranging from purified macromolecules to *in vivo* pathological tissues.

In vivo screening, with targets on endothelial cells or on tissues where the accessibility is the product of the pathological condition (e.g. tumoral cells), produces large repertoires of peptides for which the use of NGS technologies made possible a global overview. A typical setup for biomarker discovery study is the differential analysis of two repertoires corresponding to two different conditions (say, healthy vs. pathological).

Previous studies for screening of peptide libraries for organ-specific binding [3] studied sequence composition and similarity in this context. Arap et al. has shown that the distribution of the ligand-receptor pairs is non-random regarding organ specificity [4]. However, in order to prevent the problem related to the sparce count of the 20^7 possible 7-mers, they were broken into 3-mers. The work of Kolonin [5] is based on the presence of those 3-mers inside the candidate peptide-bound receptors, which identifies these targeted receptors and sites of ligands involved in receptor interaction.

In this paper we assume the working hypothesis that certain peptides mimic physiological interactions of proteins that contain a sub-sequence similar to these peptides. However, sequence similarity of peptides with a protein is a weak signal. Fine-tuned and efficient mapping of the complete peptides against the proteome of interest and an appropriate scoring function are thus the central points of our solution. In this manuscript we mainly focus on the mapping step.

The underlying assumption of our work is that small peptides mimic molecular interactions of larger proteins. This assumption is based on a number of observations:

- Proteins containing similar sub-sequences tend to themselves be homologous;
- These sub-sequences align at the same locations where homologous proteins align among themselves;
- These locations are strongly preserved by the evolutionary process (and are mostly located within functional domains);

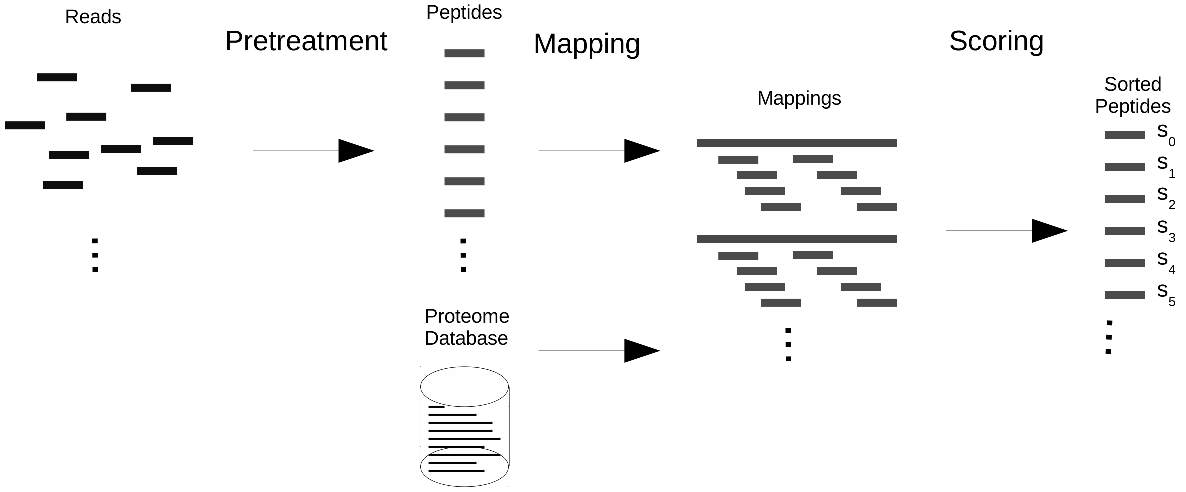
Moreover, a recent study [7] shows that most of the energy in proteins' interactions is located in short linear segments, and most of these segments bind independently of their context, which further supports our working hypothesis.

Given a peptide repertoire $R = \{r\}$ where all peptides r have the same size $s \in [7, 12]$ and proteome $P = \{p\}$, each peptide r is mapped against P, which produces a set of mapping locations L_r. This mapping is restricted to an ungapped alignment up to a certain similarity score threshold t. Once the peptides are mapped on the proteins, each peptide r is scored by integrating its mapping scores for all locations L_r combined with the scores of other peptides over the same L_r. Intuitively, it singles out those peptides that participate in those locations that are mapped by many other peptides. Consequently, the scoring function is highly sensitive to the detection of the complete set of peptides that map at a certain location L_r. This is why it is essential for a given peptide to uncover all the mappings over the given threshold t.

2 Pipeline

Our solution is implemented as a pipeline with three major steps (see figure Fig. 1 here below).

Figure 1. Peptides scoring pipeline



1. Sequenced reads are preprocessed in order to obtain the repertoire of phage displayed peptides.
2. Given this repertoire and a subject database, each peptide is mapped on each protein of the database.
3. Given the list of mappings, a scoring function sorts the peptides.

In this manuscript we concentrate on the computationally challenging step: mapping of the peptide repertoire against the proteome. We formulate this problem as simultaneous matching of multiple patterns against multiple strings and propose an algorithmically efficient solution.

3 State of the art

The central computational problem concerns mapping of R over P, which is known as the string matching problem. There exists a number of well-known algorithms for string matching [1]. They can be classified in two categories: exact and fuzzy matching.

Exact matching can itself be subdivided in two major algorithmic classes.

– Matching a simple pattern against a single string.
  These algorithms are of limited interest for our problem since we need to match multiple patterns against multiple strings. However, they provide a complexity reference point where we would match each peptide one by one against each protein. Classical examples are Boyer–Moore [8] and the text-partitioning matching [9].

– Matching multiple patterns against a single string.
  These algorithms make it possible to map R (our multiples patterns) against the proteins of the proteome one by one. A classical example of this class of algorithms is the Aho–Corasick [10].

Fuzzy matching approaches can also be classified in two major classes, both relying on the notion of metric space.

– Approximate matching
  These algorithms use the metric space to find the set of similar words. A well-known example of the underlying distance measure is the Damerau–Levenshtein distance [11] used in many spell-checkers.

– Sequence alignment
  These algorithms are particularly widespread in bioinformatics. Indeed, in this context, character mismatch or insertion/deletion are often seen as mutation points in the evolutionary process. Metric space usually relies on a substitution score matrix rather than on a fixed score in the case of mismatch. Classical examples are Smith–Waterman [12] and BLAST [13] for local alignment and Needleman–Wunch [14] for global alignment.

On one hand, our application case requires fuzzy matching, since we are looking for similarity between peptides and protein sequences. Moreover, we rely on a substitution matrix for sequence similarity. On the other hand, this is not a classical fuzzy matching problem since we are interested in the ungapped alignments. It is also important to obtain all the positions where peptides match on the proteins. Consequently, neither probabilistic (e.g, BLAST) nor best-match (e.g, Needleman–Wunsch) approaches can be used.

4 Mapping

First we preprocess the proteome $P = \{p\}$. Since we are interested in the ungapped alignments, this can be seen as fragmenting each p by a sliding window of size s, which gives us $F = \{f\}$, where f are proteic sub-sequences, that we call fragments, of size s. The problem can be now seen as matching R against F. However, this would be computationally inefficient.

To overcome this problem, we encode P in a compact way by constructing the keyword tree [15] S of F. Information on the positions L_f of each fragment and the proteins where it belongs is stored in the leaves of the tree. Another keyword tree Q is constructed for the repertoire R. To be efficient, the construction of S involves the intermediate construction of a suffix tree (over a finite alphabet, see [16]) where only the prefixes of size s of each suffix are kept.

Given a keyword tree Q (S, respectively), we denote by Q_n (S_n, respectively) any node of this tree. Each node contains a $\langle$letter, node$\rangle$ map that represents a tree edge. Given a node Q_n and a letter l, $Q_n[l]$ denotes following the corresponding edge and reaching the child of Q_n; $S_n.data$ denotes the $\langle$protein, position set$\rangle$ map.

The problem can then be formulated as mapping the two keyword trees one against another. The general matching scheme is presented in the algorithm 1 here below.

This algorithm can be trivially used to perform exact matching by replacing the scoring function by an equality comparison between l_1 and l_2. The thresholding function then becomes a placeholder present only to check the validity of the boolean score.

[1] In this section we only quote seminal works and citations are in no way exhaustive.

Algorithm 1 Recursive mapping of Q onto S

```
function MAPFUNCTION($Q_n$, $S_n$, $word$, $score$, $depth = 0$)
    $W \leftarrow$ empty set of $\langle$word, data$\rangle$ pairs
    for each letter $l_1 \in Q_n$ do
        for each letter $l_2 \in S_n$ do
            $new\_score \leftarrow$ SIMILARITYFUNCTION($l_1$, $l_2$, $score$)
            if THRESHOLDING($new\_score$, $depth$) then
                $new\_word \leftarrow$ CONCAT($word$, $l_1$)
                if $Q_n$ is a leaf node then                    ▷ $Sn$ is necessarily a leaf too
                    add ($new\_word$, $Sn.data$) pair to $W$
                else
                    $W \leftarrow W \cup$ MAPFUNCTION($Q_n[l_1]$, $S_n[l_2]$, $new\_word$, $new\_score$, $depth + 1$)
                end if
            end if
        end for
    end for
    return $W$
end function
```

4.1 Similarity score

Since our mapping is ungapped, in order to compute the *similarity* between a peptide r and a fragment f, it is sufficient to compute the letter-wise substitution cost based on a substitution matrix M, namely:

$$score_{r/f} = \frac{2\sum_{l_r \in r, l_f \in f} M(l_r, l_f)}{\sum_{l_r \in r} M(l_r, l_r) \sum_{l_f \in f} M(l_f, l_f)}.$$

Using this formula, similarity of two exact same words is 1.0 and the similarity score drops with each substitution involved.

Algorithm 1 is recursive, consequently, both similarity score computation and the thresholding have to respect the recursivity.

- For the similarity this implies to maintain through the recursive call both the numerator and the denominator independently. And it is within the thresholding function that the fraction is computed.
- For the thresholding this implies that the threshold value has to be defined as function of the number of letters that have been processed, which means that it is different for each tree depth. Were this computation done naively, it would require to check for the threshold only when all of the letters have been processed. Indeed, the first processed letters may drop the score below the threshold even if the similarity score can increase after that. In the recursive solution we can compute at each tree depth, given the word size s and the maximal value in the substitution matrix $\max(M)$, the maximal similarity score with respect to the letters that have already been processed.

Resulting functions can be seen in the algorithm 2.

Algorithm 2 Recursive scoring and thresholding

```
function SIMILARITYFUNCTION($l_1$, $l_2$, $score$: ($num$, $den$))
    $n \leftarrow 2M(l_1, l_2)$
    $d \leftarrow M(l_1, l_1) + M(l_2, l_2)$
    return pair ($num + n$, $den + d$)
end function

function THRESHOLDING($score$: ($num$, $den$, $depth$))
    $k \leftarrow 2(s - depth - 1)\max(M)$
    return $(num + k)/(den + k) > t$
end function
```

References

[1] G. P. Smith, Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315-1317, 1985.

[2] D. Wacker, C. Wang, V. Katritch, G. W. Han, X. Huang, E. Vardy, J. D. McCorvy, Y. Jiang, M. Chu, F. Y. Siu, W. Liu and H. E. Xu, V. Cherezov, B. L. Roth and R. C. Stevens, Structural Features for Functional Selectivity at Serotonin Receptors. *Science*, doi: 10.1126/science.1232808, 2013.

[3] R. Pasqualini and E. Ruoslahti, Organ targeting in vivo using phage display peptide libraries. *Nature*, 380(6572):364-366, 1996.

[4] W. Arap, M. G. Kolonin, M. Trepel, J. Lahdenranta, M. Cardó-Vila, R. J. Giordano, P. J. Mintz, P. U. Ardelt, V. J. Yao, C. I. Vidal, L. Chen, A. Flamm, H. Valtanen, L. M. Weavind, M. E. Hicks, R. E. Pollock, G. H. Botz, C. D. Bucana, E. Koivunen, D. Cahill, P. Troncoso, K. A. Baggerly, R. D. Pentz, K. A. Do, C. J. Logothetis and R. Pasqualini, Steps toward mapping the human vasculature by phage display. *Nature Medicine*, 8(2):121-127, 2002.

[5] M. G. Kolonin, J. Sun, K. A. Do, C. I. Vidal, Y. Ji, K. A. Baggerly, R. Pasqualini and W. Arap, Synchronous selection of homing peptides for multiple tissues by in vivo phage display. *The FASEB Journal*, 20(7):979-981, 2006.

[6] L. G. León-Novelo, P. Müller, W. Arap, M. Kolonin, J. Sun, R. Pasqualini and K. A. Do, Semiparametric Bayesian Inference for Phage Display Data. *Biometrics*, doi: 10.1111/j.1541-0420.2012.01817.x, 2013.

[7] N. London, B. Raveh, D. Movshovitz-Attias and O. Schueler-Furman, Can Self-Inhibitory Peptides be Derived from the Interfaces of Globular Protein-Protein Interactions?. *Proteins*, 78(15):3140-3149, 2010.

[8] R. S. Boyer and J. S. Moore, A fast string searching algorithm. *Communications of the ACM*, 20(10):762-772, 1977.

[9] S. Kim, A new string-pattern matching algorithm using partitioning and hashing efficiently. *Journal of experimental algorithmics*, 4, 1999.

[10] A. V. Aho and M. J. Corasick, Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333-340, 1975.

[11] F. J. Damerau, A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171-176, 1964.

[12] T. F. Smith and M. S. Waterman, Identification of common molecular subsequences. *Journal of menucular biology*, 147(1):195-197, 1981.

[13] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, Eugene W. and D. J. Lipman, Basic local alignment search tool. *Journal of molecular biology*, 215(3):403-410, 1990.

[14] S. B. Needleman and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443-453, 1970.

[15] D. Gusfield, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, Cambridge University Press, 1997.

[16] E. Ukkonen, On-line construction of suffix trees. *Algorithmica*, 14(3):249-260, 1995.

Adaptive Smith-Waterman Residue Match Seeding for Protein Structural Alignment

Christopher M. TOPHAM[1-3], Mickaël ROUQUIER[1-3], Nathalie TARRAT[1-3] and Isabelle ANDRE[1-3]

[1] UNIVERSITE DE TOULOUSE, INSA, UPS, INP, 135 Avenue de Rangueil, F-31077, Toulouse, France
[2] CNRS, UMR5504, F-31400 Toulouse, France
[3] INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, F-31400, Toulouse, France
{christopher.topham, isabelle.andre}@insa-toulouse.fr

Keywords Amino acid residue physical environment, ligand contact, peptide fragment, rigid-body superposition, structural overlap, distance metric, SCOP, ASTRAL.

Introduction

Structural comparison and alignment play a pivotal role in the understanding of evolutionary and functional relations in proteins, providing the material basis of both hierarchical protein classification systems and continuous descriptions of protein fold space. Methods for protein structural alignment and superposition of equivalenced amino acid residue positions are employed in a wide range of additional practical applications, including comparative protein modelling, ligand binding-site identification, protein-ligand docking, small molecule overlay for structure-based ligand design, protein functional annotation and design. Structure-based alignment is also used to improve the sequence alignment quality of remote homologues, and in the seeding and amplification of multiple sequence alignments in which pairwise sequence identity similarities fall below threshold reliability limits of alignment methods that consider residue identity alone.

The demand for accurate automated classification and functional annotation of a growing number of available experimental structures, many bearing the fruits of structural genomics initiatives, and their accompanying use in a diverse range of design applications, has led to the development of a large number of algorithms for protein structural comparison and alignment. Proposed methods differ widely in terms of the dimensionality used to represent the protein residue string, the scoring (objective) function for structural similarity and the heuristic algorithm used to optimise the objective function. Recent studies of comparative performance highlight a need to improve the geometric quality and consistency of computed structural alignments [1-5]. The principal difficulties concern the accommodation of natural structural variation and plasticity in protein alignment, complicated by a necessity for concordant measures of structural similarity.

Here, we describe the POLYFIT adaptive rigid-body algorithm for automated global pairwise and multiple protein structural alignment.

Results

Smith-Waterman local alignment [6] is used by POLYFIT to establish a robust set of seed equivalences that are extended using Needleman-Wunsch global dynamic programming techniques. Structural and functional interaction constraints provided by evolution are encoded as 1-D residue physical environment strings for the rapid alignment of highly structurally overlapped protein pairs, spanning the Doolittle twilight zone (from 20% to 30% sequence identity) and beyond. Amino acid residue physical descriptors include secondary structure, solvent accessibility, hydrogen bonding, disulphide bridging, aromatic occluded contact surface, metal ion co-ordination and small molecule ligand interaction.

Local structure alignment of more distantly related protein pairs, generally with < 20% shared sequence identity, is carried out using 3-D rigid-body conformational matching of 15-residue fragments, with allowance in the heuristic for less stringent conformational matching of ligand contact, disulphide bridge and *cis*-peptide correspondences. Structural similarity and alignment difficulty are assessed using the 3.5 Å overlap parameter [7], quasi-identical to the pairwise fractional topological weighted component of a distance metric [8-9] for structure-based phylogenetic clustering [10-11]. Protein structural plasticity is accommodated through the stepped adjustment of a single empirical distance parameter value in the calculation of the Smith-Waterman dynamic programming matrix.

Pairwise alignment accuracy has been bench-marked against that of ten widely used aligners on the Sippl and Wiederstein [12] set of difficult pairwise structure alignment problems, and more extensively against that of Matt [13], SALIGN [7] and MUSTANG [14] in pairwise and multiple structural alignments of 8446 protein domains belonging to 1789 SCOP families [15] with low shared sequence identity in the ASTRAL [16] 40% compendium. The 3.5 Å structural overlap parameter is used to assess alignment quality [5]. The results demonstrate that POLYFIT performs equally well or better than other aligners in a large majority of test cases, and remains computationally efficient in the alignment of distantly related protein structures.

POLYFIT is available on the web-server at http://polyfit.insa-toulouse.fr. Users are required to upload pre-extracted protein domains as heavy-atom PDB-formatted co-ordinate files conforming to version 3 of the RCSB PDB chemical component dictionary (http://www.wwpdb.org/ccd.html). Pre-processed residue physical environment-annotated C^{α} atomic co-ordinate sets are available for alignment on the web server for 19,052 protein domains in the ASTRAL 40% and 95% sub-sets of the January 2013 SCOP v1.75B release.

Acknowledgement

This work was supported in part by the French ANR national research agency.

References

[1] H. Hasegawa and L. Holm, Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.*, 19: 341-348, 2009.

[2] R. Kolodny, P. Koehl and M. Levitt, Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, 346: 1173-1188, 2005.

[3] C. Kim and B. Lee, Accuracy of structure-based sequence alignment of automatic methods. *BMC Bioinf.*, 8: 355, 2007.

[4] M.I. Sadowski and W.R. Taylor, Evolutionary inaccuracy of pairwise structural alignments. *Bioinformatics*, 28: 1209-1215, 2012.

[5] A.W. Slater, J.I. Castellanos, M.J. Sippl and F. Melo, Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments. *Bioinformatics*, 29: 47-53, 2013.

[6] T.F. Smith and M.S. Waterman, Identification of common molecular subsequences. *J. Mol. Biol.*, 147: 195-197, 1981.

[7] M.S. Madhusudhan, B.M. Webb, M.A. Marti-Renom, N. Eswar and A. Sali, Alignment of multiple protein structures based on sequence and structure features. *Protein Eng. Des.*, 22: 569-574, 2009.

[8] M.S. Johnson, A. Šali and T.L. Blundell, Phylogenetic relationships from three-dimensional protein structures. *Methods Enzymol.*, 183: 670-690, 1990.

[9] M.S. Johnson, M.J. Sutcliffe and T.L. Blundell, Molecular anatomy: phyletic relationships from three-dimensional protein structures. *J. Mol. Evol.*, 30: 43-59, 1990.

[10] G. Argawal, M., Rajavel, B. Gopal and N. Srinivasan, Structure-based phylogeny as a diagnostic for functional characterization of proteins with a cupin fold. *PLos One*, 4: e5736, 2009.

[11] L.S. Pidigu, K. Maity, K. Ramaswamy, N. Suriola and K. Sugana, Analysis of proteins with the 'hot dog' fold: prediction of function and identification of catalytic residues in hypothetical proteins. *BMC Struct. Biol.*, 9: 37, 2009.

[12] M.J. Sippl and M. Wiederstein, A note on difficult structure alignment problems. *Bioinformatics*, 24: 426-427, 2008.

[13] M. Menke, B. Berger and L. Cowen, Matt: local flexibility aids protein multiple structural alignment. *PLoS Comp. Biol.*, 4: e10, 2008.

[14] A.S. Konagurthu, J.C. Whisstock, J.A. Irving, P.J. Stuckey and A.M. Lesk, MUSTANG: a multiple structural alignment algorithm. *Proteins: Struct. Funct. Bioinf.*, 64: 559-574, 2006.

[15] A. Andreeva, D. Howarth, J.-M. Chandonia, S.E. Brenner, T.J.P. Hubbard, C. Chothia and A.G. Murzin, Data growth and its impact on the SCOP database. *Nucleic Acids Res.*, 36: D419-D425, 2008.

[16] J.-M. Chandonia, G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt and S.E. Brenner, The ASTRAL compendium in 2004. *Nucleic Acids Res.*, 32: D189-D192, 2004.

miRNAs Networks in Human Embryonic Stem Cells

Ricky BHAJUN[1,2,3], Laurent GUYON[1,2,3], Eric SULPICE[1,2,3], Stéphanie COMBE[1,2,3], Lay Teng ANG[4],
Christian LAJAUNIE[5], Bing LIM[4] and Xavier GIDROL[1,2,3]

[1] LABORATOIRE BGE, iRTSV, CEA, 17 rue des Martyrs, 38054 Grenoble cedex 9, France

[2] INSERM, U1038, 38054 Grenoble cedex 9, France

[3] UJF, F-38000, Grenoble 1, France

{ricky.bhajun, laurent.guyon, eric.sulpice, stephanie.combe,
xavier.gidrol}@cea.fr

[4] STEM CELL AND DEVELOPMENTAL BIOLOGY, GIS, 60, Biopolis Street, 138672, Singapore

limb1@gis.a-star.edu.sg, anglayteng@gmail.com

[5] Mines-ParisTech, Institut Curie, INSERM U900, 26 rue d'Ulm, 75248 Paris cedex 05, France

christian.lajaunie@mines-paristech.fr

Keywords microRNAs, gene regulation, human embryonic stem cells (hESCs), differentiation, totipotency, networks, meta-analysis.

Isolated from the inner cell mass of blastocyst, embryonic stem cells (ESCs) are a particular type of cells which have the ability to self-renew and differentiate into specialized cell types. These two abilities are fine-tuned and controlled by transcription factors (TFs) [1] but it is known that some small non coding RNAs (miRNAs) also plays a role in the regulation of stem cell state [2]. Many works have been conducted trying to decipher which transcription factors and miRNAs regulate stem cells fate. Noteworthy is Chia *et al.*'s siRNA genome-wide functional screen where they identified a number of genes, including transcription factors, involved in the maintenance of totipotency of hESCs [3].

We used data from an unpublished genome-wide functional screen that we performed in our teams with a set of Locked Nucleic Acid (LNA) inhibiting miRNAs in hESCs. We used a cell line that constitutively express an OCT4-GFP reporter construction as described in Chia *et al.* [3]. Thus, lowering of GFP fluorescence after transfection (compared to controls) could be linked to a loss of totipotency state for the cells whereas an increase or no difference of GFP fluorescence was associated with totipotency. We also used a novel cell-to-cell scoring method (called Gscore, see JOBIM abstract of Guyon *et al.*) to select hits, that is, miRNAs implicated in the maintenance of totipotency in hESCs. We then predicted gene targets for the best miRNAs hits of the screen by DIANA-microT v3 [4] and compared these results with Chia *et al.*'s screening hits (Figure 1.A.). One drawback of microRNAs targets prediction algorithms is their false positive and negative rates which are hard to estimate. DIANA-microT was selected for its good precision in [5] compared to other software, namely 66%, and its score ability to correlate with gene expression.

To build our network, we used the predictions from DIANA-microT v3 on a miRnome-wide scale and used Cytoscape [6] as a visualization tool. We inferred a link between two miRNAs only if fifty percent of targets for one miRNA were also predicted as potential targets for a second, thus specifying the possibility that these two miRNAs are implied in the same biological pathway by acting on the similar set of genes, as described in Shalgi *et al.* [7]. The network is composed of 555 nodes (miRNAs) and 2911 edges (at least 50% shared targets) and exhibits a scale free behavior (Figure 1.B.). Without *a priori* knowledge, clusters of genes that were well described in the literature were retrieved in the network. We also mapped the results obtained from the LNA screen on the network to determine which miRNAs in this network were involved in hESCs fate.

Interestingly, the average of the miRNAs' absolute Gscores for the cluster shown in figure 1.A.α. gives a score that is significantly higher compared to the score calculated for microRNAs chosen by chance in the network (pValue 0.02) demonstrating that the whole cluster is statistically highly involved in the maintenance of the totipotency state of the cell. Even though some miRNAs are found with Gscore close to 0 in this cluster, these miRNAs might still have a role in the totipotency state: we could in fact simply "miss" the phenotype of interest (e.g. the time when the images were taken). Indeed, hsa-miR-302a and has-miR-371, both belonging to this cluster, show low Gscores but are in fact already known to be involved in the maintenance of totipotency [8]. As such, and for this example, the whole cluster should be prioritized for further experiment.

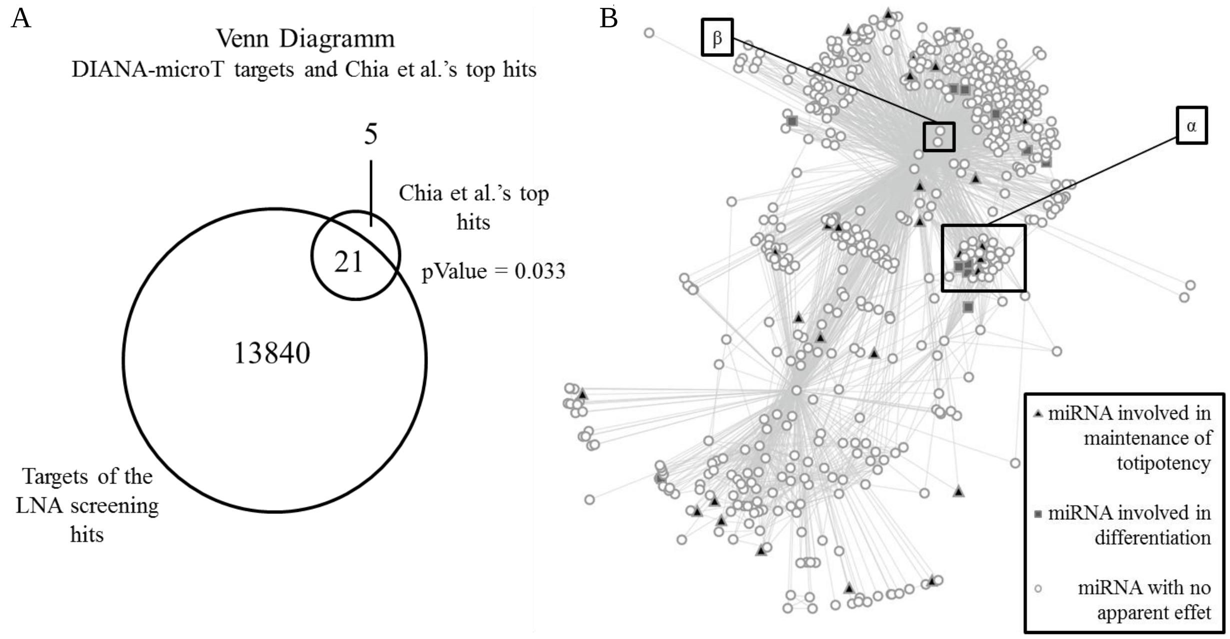

A
Venn Diagramm
DIANA-microT targets and Chia et al.'s top hits
5
Chia et al.'s top hits
21
pValue = 0.033
13840
Targets of the LNA screening hits
B
β
α
▲ miRNA involved in maintenance of totipotency
■ miRNA involved in differentiation
○ miRNA with no apparent effet

Figure 1. **A.** Coverage of DIANA-microT predictions and Chia *et al.*'s top hits. **B.** miRNAs network from DIANA-microT (colored by the LNA screen). α) Cluster of miRNAs well known to be involved in hESCs maintenance and differentiation. β) The two biggest miRNAs hubs.

References

[1] H. Niwa, How is pluripotency determined and maintained? 134:635‑646, 2007.

[2] A. Marson, S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, J. M. Calabrese, L. M. Dennis, T. L. Volkert, S. Gupta, J. Love, N. Hannett, P. A. Sharp, D. P. Bartel, R. Jaenisch, et R. A. Young, Connecting microRNA Genes to the Core Transcriptional Regulatory Circuitry of Embryonic Stem Cells, *Cell*, 134:521‑533, 2008.

[3] N.-Y. Chia, Y.-S. Chan, B. Feng, X. Lu, Y. L. Orlov, D. Moreau, P. Kumar, L. Yang, J. Jiang, M.-S. Lau, M. Huss, B.-S. Soh, P. Kraus, P. Li, T. Lufkin, B. Lim, N. D. Clarke, F. Bard, et H.-H. Ng, « A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity », *Nature*, 468:316‑320, 2010.

[4] M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, et A. G. Hatzigeorgiou, DIANA-microT web server: elucidating microRNA functions through target prediction, *Nucl. Acids Res.*, 37:W273‑W276, 2009.

[5] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, et N. Rajewsky, Widespread changes in protein synthesis induced by microRNAs, *Nature*, 455:58‑63, 2008.

[6] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, et T. Ideker, Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks, *Genome Res*, 13:2498‑2504, 2003.

[7] R. Shalgi, D. Lieber, M. Oren, et Y. Pilpel, Global and Local Architecture of the Mammalian microRNA–Transcription Factor Regulatory Network, 3:e131, 2007.

[8] B. Stadler, I. Ivanovska, K. Mehta, S. Song, A. Nelson, Y. Tan, J. Mathieu, C. Darby, C. A. Blau, C. Ware, G. Peters, D. G. Miller, L. Shen, M. A. Cleary, et H. Ruohola-Baker, Characterization of microRNAs involved in embryonic stem cell states, *Stem Cells Dev.*, 19:935‑950, 2010.

Gscore, a Robust Cell-by-cell Score Enhances Sensitivity and Specificity for High Content Screening Hit Discovery

Laurent GUYON[1,2,3], Christian LAJAUNIE[4,5,6], Frédéric FER[1,2,3], Ricky BHAJUN[1,2,3], Mélissa MARY[1,2,3], Eric SULPICE[1,2,3], Stéphanie COMBE[1,2,3], Patricia OBEID[1,2,3], Jean-Philippe VERT[4,5,6], Xavier GIDROL[1,2,3]

[1] LABORATOIRE BGE, iRTSV, CEA, 17 rue des Martyrs, F-38054 Grenoble cedex 9, France

[2] INSERM, U1038, F-38054 Grenoble cedex 9, France

[3] UJF, F-38000, Grenoble 1, France

{laurent.guyon, ricky.bhajun, eric.sulpice, stephanie.combe, patricia.obeid, xavier.gidrol}@cea.fr

[4] Centre for Computational Biology - CBIO, Mines ParisTech, 35 rue Saint-Honoré, Fontainebleau, F-77300 France

[5] Institut Curie, 26 rue d'Ulm, Paris, F-75248 France

[6] INSERM, U900, Paris, F-75248 France

{christian.lajaunie, Jean-Philippe.Vert}@mines-paristech.fr

Keywords High Content Screening (HCS), scoring, hit finding, False Positive, Area Under ROC Curve (AUC).

1 Introduction

High Content Screening (HCS) has enabled great advances both in oncology and biology. It consists in visualizing phenotypes modification of cells after perturbation. This perturbation is generally done either by chemical compounds or RNA interference (including silencing RNAs), in a highly parallel manner (in 384 well-plates for example).

HCS experiments produce tremendous amount of data, typically tables of millions of rows and tens of columns; each row corresponding to a cell which phenotype is characterized in the different columns. After analysis, hits, which are the biggest modifiers of the cell phenotype, are extracted. However, the technique still needs improvements to lower false positive rate, enhance sensitivity and help hit and pathway discovery, particularly in physiological conditions such as low transfection of RNAi with hard to transfect cells (stem cells, primary cells and suspension cells), which can generate numerous artifacts.

2 Gscore, a Cell-by-cell Score for Hit Detection

To face these issues, we have developed Gscore, a novel scoring function based on robust analysis at single cell level. Variability due to the number of cells is also taken into account to provide an optimal aggregation of scores among replicates. To test performance of the scoring functions, we performed both simulated and experimental cell-by-cell screening with various and fixed parameters. In these two dedicated screens, conditions are controlled, allowing the comparison of different scoring functions, including Gscore, Zscore, Robust Zscore, LZscore and SSMD [1], and another cell-by-cell scoring function adapted from GSEA [2]. As positive (that is effective) and negative (ineffective) perturbations are known, sensitivity and specificity of each scoring function are estimated. Furthermore, Area Under the ROC Curve (AUC) is calculated and used to judge scoring function performance.

Using simulated data, the efficiency of this new scoring method was demonstrated in various conditions. The Gscore performs always better in term of sensitivity and selectivity than the widely used Zscore formula, especially in difficult screening conditions. For instance, highly variable signal analyzed with Zscore will enable hit selection only if more than 90% of the cells are transfected (meaning affected by the perturbation), while Gscore performs similarly with only 35% of transfection efficiency (figure 1). Gscore even supports the development of new screening techniques that lower the minimum number of cells and of replicates per condition, which is of particular importance when it is difficult to cultivate cells such as for primary cancer cells obtained from patients.

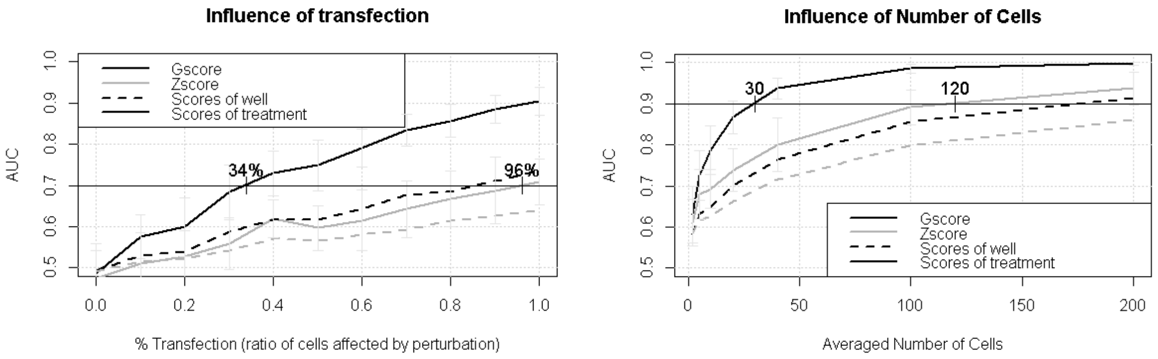


Figure 1. Performance of Gscore and Zscore on simulated screening data. Left. AUC as a function of transfection efficiency (percentage of cells that are affected by the treatment). Right. AUC as a function of the number of cells.

A controlled experiment was then performed to assess the performance of Gscore in real screening conditions, using a 384 well-plate. In this experiment done on fluorescent cells expressing GFP, only controls were used: negative control with siRNA known to have no effect, and positive control with siGFP known to decrease the fluorescence of the cells. Each control was used at different concentration to judge the sensitivity of the scores. Figure 2 shows that Gscore outperforms Zscore in real conditions, the effect being more evident for low cell number.

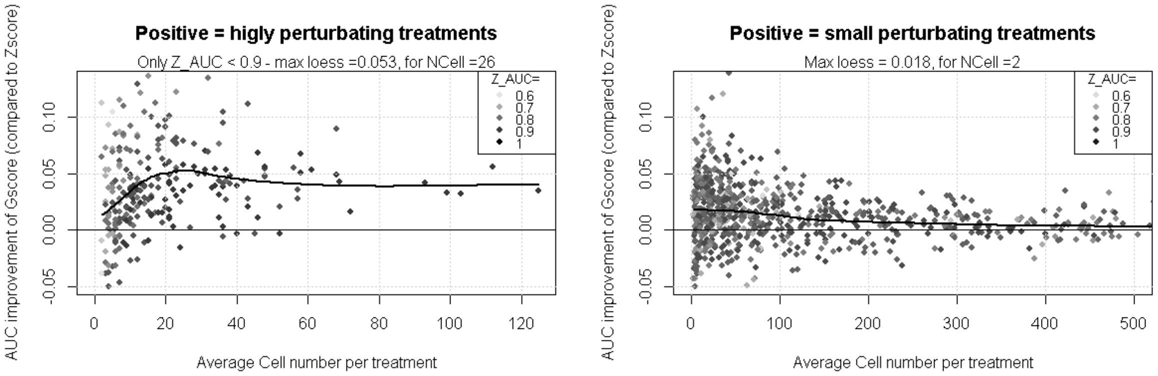


Figure 2. Performance of Gscore and Zscore on simulated screening data. Left. AUC as a function of transfection efficiency (percentage of cells that are affected by the treatment). Right. AUC as a function of the number of cells.

References

[1] X. D. Zhang, “Illustration of SSMD, z score, SSMD*, z* score, and t statistic for hit selection in RNAi high-throughput screens.,” *Journal of biomolecular screening*, vol. 16, no. 7, pp. 775–85, Aug. 2011.

[2] B. Knapp, I. Rebhan, A. Kumar, P. Matula, N. a Kiani, M. Binder, H. Erfle, K. Rohr, R. Eils, R. Bartenschlager, and L. Kaderali, “Normalizing for individual cell population context in the analysis of high-content cellular screens.,” *BMC bioinformatics*, vol. 12, no. 1, p. 485, Jan. 2011.

Some Recent Progress on Generating Symmetric Protein Complexes Using Spherical Polar Fourier Docking Correlations

David W. RITCHIE[1] and Sergei GRUDININ[2]

[1] Inria Nancy – Grand Est, LORIA, UMR7503, 615 Rue du Jardin Botanique, 54600 Villers-lès-Nancy, Cedex, France
Dave.Ritchie@inria.fr

[2] Inria Grenoble – Rhone-Alpes, UMR5224, B.P. 53, 38041 Grenoble, Cedex 9, France
Sergei.Grudinin@inria.fr

Keywords protein complexes, point group symmetry, symmetry docking, Fourier correlation.

1 Introduction

Many of the protein complexes in the protein Data bank (PDB) are symmetric homo-oligomers. Table 1 summarises the number of symmetric complexes reported by the 3D-Complex database [1]. This shows that C_2 homo-dimers comprise the majority of known homo-oligomers. However, many complexes have higher order rotational symmetry (i.e. $C_{n>2}$), and a good number have multiple rotational symmetry axes, namely those with dihedral (D_n), tetrahedral (T), octahedral (O), and octahedral (I) point group symmetries. Although symmetrical complexes are often solved directly by X-ray crystallography, it would still be very useful to be able to predict whether or not a given monomer might self-assemble into a symmetrical structure.

n	2	3	4	5	6	7	8
Cn	8740	992	223	107	76	29	5
Dn	2111	585	173	46	20	23	6

Table 1. The number of Cn and Dn complexes in the 3D-Complex database. 3D-Complex also reports 86 tetrahedral, 47 octahedral, and 6 icosahedral complexes (the 3D-Complex database excludes all viral structures).

This article briefly introduces a new point group symmetry docking algorithm that we are developing. In the last few years, several protein-protein docking programs have been adapted to apply various geometric filtering constraints to select approximately symmetrical pair-wise docking orientations [2,3,4]. However, to our knowledge, there does not yet exist an algorithm which can automatically generate perfectly symmetrical protein complexes for arbitrary point group symmetry types.

2 Methods

The usual convention for the C_n and D_n symmetry groups is to let z be the principal symmetry axis, so that applying successive rotations of $\omega = 2\pi/n$ about this axis leaves an indistinguishable arrangement of objects. This is illustrated in Fig. 1. The D_n groups have n additional two-fold rotational axes perpendicular to the principal symmetry axis. In order to generate protein complexes which automatically satisfy such symmetries, it is useful to introduce the notion of a "docking equation" in which the notation

$$A(\underline{x}) \longleftrightarrow B(\underline{x}) \tag{1}$$

represents an interaction between proteins A and B in 3D space. It is also useful introduce the operators $\hat{T}(x, y, z)$ and $\hat{R}(\alpha, \beta, \gamma)$ which represent the actions of translating an object by an amount (x, y, z) and rotating it according to the three Euler rotation angles (α, β, γ). Then, guided by Fig. 1, and assuming that we start with two identical monomers at the origin, we can write down a C_n docking equation for the two monomers as

$$\hat{T}(0, y, 0)\hat{R}(\alpha, \beta, \gamma)A(\underline{x}) \longleftrightarrow \hat{R}(0, 0, \omega)\hat{T}(0, y, 0)\hat{R}(\alpha, \beta, \gamma)B(\underline{x}). \tag{2}$$

In other words, proteins A and B are each rotated at the origin by a certain amount and then translated along the positive y axis, and protein B is finally rotated into its symmetry-related position with respect to A. A similar docking equation can be written to describe D_n complexes. Of course, we do not know in advance how to rotate and translate the proteins to make a symmetric complex. Therefore, we perform a series one-dimensional fast Fourier transform (FFT) correlation searches using the *Hex* spherical polar Fourier docking algorithm [5] to determine the four parameters $(y, \alpha, \beta, \gamma)$. We can then build the complex by substituting these parameters into Equation 2 and by repeatedly applying $\hat{R}(0, 0, \omega)$ to the solved position of the B monomer.

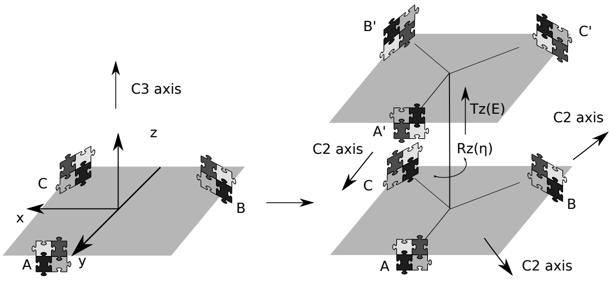


Figure 1. Illustrations of the C_3 and D_3 point group symmetries. For C_3 symmetry (left), applying a rotation of $2\pi/3$ about the principal axis gives an indistinguishable arrangement of objects. A D_3 system (right) may be constructed from two planar C_3 systems. In addition to the principal C_3 axis, D_3 also has three perpendicular two-fold axes, as shown.

3 Results and Conclusion

Fig. 2 shows some examples of C_n complexes that we have generated starting from a complex between Kallikrein A and bovine pancreatic trypsin inhibitor (PDB code 2KAI). Each complex is perfectly symmetrical, although due to the the soft docking function in *Hex* it is possible that some interfaces might contain minor steric clashes. The calculation for each structure takes about 20 seconds on a modern workstation. Despite the fact that the initial complex is very unlikely to form symmetrical oligomers, our algorithm has produced some remarkably plausible structures. We are currently extending the technique to build D_n, T, O, and I symmetries in a similar way, and we are collecting a benchmark of examples with which to evaluate the overall approach.

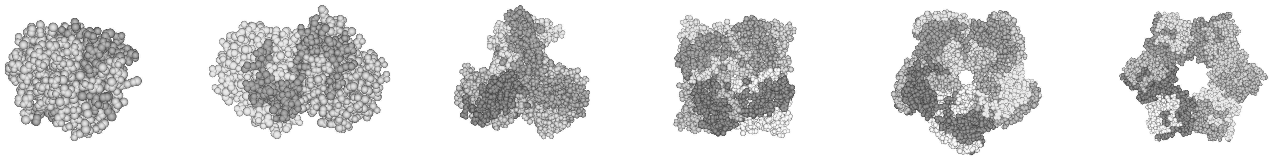


Figure 2. Some artificial symmetry complexes generated by our approach. Looking from left to right, the first complex is the starting structure (PDB code 2KAI). The subsequent structures are an artificial C_2 dimer, C_3 trimer, C_4 tetramer, C_5 pentamer, and C_6 hexamer, each generated from the original complex.

Acknowledgements

This work was supported by the ANR (grant numbers ANR-MONU-006-01 and ANR-MONU-006-02).

References

[1] E. D. Levy, E. Boeri Erba, C. V. Robinson, and S. A. Teichmann. Assembly reflects evolution of protein complexes. *Nature*, 453:1262–1266, 2008.

[2] S. R. Comeau and C. J. Camacho. Predicting oligomeric assemblies: N-mers a primer. *Journal of Structural Biology*, 150:233–244, 2004.

[3] B. Pierce, W. Tong, and Z. Weng. M-ZDOCK: a grid-based approach for C$_n$ symmetric multimer docking. *Bioinformatics*, 21:1472–1478, 2005.

[4] D. Schneidman-Duhovny, Y Inbar, R Nussinov, and H.J. Wolfson. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Research*, 33:W363–W367, 2005.

[5] D. W. Ritchie and G. J. L. Kemp. Protein docking using spherical polar Fourier correlations. *Proteins: Structure, Function, Genetics*, 39(2):178–194, 2000.

Why you should care about protein size when benchmarking protein-protein interface predictions

Juliette MARTIN[1]

[1] BASES MOLECULAIRES ET STRUCTURALES DES SYSTEMES INFECTIEUX, UMR5086 CNRS/Université
Lyon, 7 passage du Vercors, 69367 , Lyon, Cedex 07, France
juliette.martin@ibcp.fr

Keywords protein-protein interaction, protein interface prediction, size bias, over-estimation, performance

In this poster, we investigate the effect of the size bias in the prediction of protein-protein interface. The size bias originates in the fact that the proportion of interface residues varies with the size of a protein: small protein have more interface residues, in proportion. Because of that, any prediction method that does not correct for the size effect can be affected by a bias in the assessment when it is done on a data set containing both small and large proteins.

Using simulated prediction scores, we explain and quantify the size bias using the Docking Benchmark version 4.0 as a test data set. We then show how to detect and correct the scores, in order to remove the size bias, using the results of a real prediction method. All the eight methods tested so far suffer from an over-estimation of their performance when assessed on the Docking Benchmark 4.0. Their corrected AUC (Area Under the ROC curve, a performance index that varies between 0.5 for random to 1 for perfect predictor) are 0.02 to 0.05 lower than before the correction.

This indicates that the size effect is often overlooked and that we all should systematically track it and correct it.

The evolution and biosynthetic potential of secondary metabolites in cyanobacteria

Alexandra CALTEAU[1], David P. FEWER[2], and Muriel GUGGER[3]

[1] CEA, DSV, Institut de Génomique (IG), Genoscope and CNRS UMR 8030, Laboratoire d′Analyses Bioinformatiques en Génomique et Métabolisme, 91057 Evry, France
acalteau@genocope.cns.fr

[2] Division of Microbiology, Department of Food and Environmental Sciences, University of Helsinki, FIN-00014, Helsinki, Finland
david.fewer@helsinki.fi

[3] Institut Pasteur, Collection des Cyanobactéries, 75724 Paris Cedex 15, France
muriel.gugger@pasteur.fr

Keywords: Cyanobacteria, secondary metabolites, NRPS, PKS, evolution

Polyketide synthase (PKS), and non-ribosomal peptide synthetase (NRPS) are modular biosynthetic enzymes responsible for the production of structurally diverse and biologically important natural products [1]. These non-ribosomal pathways allow the production of small but complex molecules containing a large range of proteinogenic and non-proteinogenic amino acids. Cyanobacteria are prolific producers of secondary metabolites. Hundreds of bioactive compounds with diverse chemical structures have been described from this phylum [2].

The Cyanobacteria were recently shown, through a phylum-wide genome investigation, to encode a tremendous number of gene clusters devoted to the biosynthesis of secondary metabolites [3]. Large-scale bioinformatics analyses were performed on the 126 genomes of the CyanoGEBA dataset [3] covering a set of phylogenetically diverse cyanobacteria with a great breadth of morphologies and ecologies. We identified 454 NRPS/PKS clusters, which were present in 71% of the genomes investigated. We estimated that non-ribosomal gene clusters occupy up to 5.5% of the genome in some strains.

The huge chemical diversity of cyanobacterial compounds is reflected at the genomic level with a huge diversity of pathway synthesis. Most of the clusters await further characterization; however, the potential for the production of known toxins and compounds can already be predicted. We also identified an unexpected potential for siderophore biosynthesis.

This first phylum wide study also highlights the complex history of secondary metabolites production in Cyanobacteria. The genomic content in NRPS/PKS gene clusters varies greatly along the cyanobacterial phylogenetic tree suggesting different evolutionary scenarios at the phylum level. Genomic comparison suggests a complex evolution of NRPS/PKS clusters with gene shuffling and horizontal gene transfer events.

References

[1] JL Meier, MD Burkart, The chemical biology of modular biosynthetic enzymes. Chem Soc Rev., 38(7):2012-2045, 2009

[2] M. Welker, H. von Döhren, Cyanobacterial peptides - nature's own combinatorial biosynthesis, *FEMS Microbiol Rev.,* 30(4):530-563, 2006

[3] PM Shih, D Wu, A Latifi, SD Axen, DP Fewer, E Talla, A Calteau, F Cai, N Tandeau de Marsac, R Rippka, M Herdman, K Sivonen, T Coursin, T Laurent, L Goodwin, M Nolan, KW Davenport, CS Han, EM Rubin, JA Eisen, T Woyke, M Gugger, CA Kerfeld, Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci U S A.,* 15;110(3):1053-1058, 2013

Nucleosome positioning in *Paramecium tetraurelia*

Marion WEIMAN[1], Mélody MATELOT[2], Olivier ARNAIZ[1], Cédric VAILLANT[3], Frédéric GUERIN[2],
Yves d'AUBENTON-CARAFA[1], Mireille BÉTERMIER[1], Sandra DUHARCOURT[2], Claude THERMES[1] and
Chun-Long CHEN[1]

[1]CENTRE DE GENETIQUE MOLECULAIRE, UPR CNRS 3404, 91198 Gif-sur-Yvette, France
{weiman, arnaiz, daubenton, betermier, thermes, chen}@cgm.cnrs-gif.fr
[2]INSTITUT JACQUES MONOD, CNRS, UMR 7592, UNIVERSITE PARIS DIDEROT, SORBONNE PARIS CITE, 75205 Paris, France
{matelot, guerin, duharcourt}@ijm.univ-paris-diderot.fr
[3]LABORATOIRE DE PHYSIQUE, ECOLE NORMALE SUPERIEURE DE LYON, CNRS, 69364 Lyon, France
{cedric.vaillant}@ens-lyon.fr

Keywords Nucleosome, mRNA processing, intron recognition, evolution.

1 Introduction

Nucleosomes are formed by approximately 147 base pairs (bp) of genomic DNA wrapped around octamers of histone proteins. Nucleosome positioning can affect DNA accessibility to the nuclear environment and as such plays an essential role in the regulation of cellular processes, including gene transcription regulation and mRNA processing (reviewed in [1]). Most eukaryotic genes are interrupted by introns that must be accurately removed from pre-messenger RNAs to produce translatable mRNAs. In multicellular eukaryotes, long introns are recognized through exon definition. Recent studies have shown that nucleosomes stably positioned along exons seem to contribute to define the exon-intron architecture, possibly pointing to a function in exon definition [2, 3]. By contrast, short introns are recognized through intron definition. With an average length of 25 nucleotides (nt), introns of *Paramecium tetraurelia*, a unicellular ciliate, are among the shortest reported in eukaryotes [4]. The large number of introns (> 90,000 predicted, 2.3 introns/gene) are associated with weak splice sites (only the first and last three intron bases are highly constrained). The mechanisms specifying the correct recognition of these tiny introns remain poorly understood and the role of nucleosome positioning in *Paramecium* intron recognition has not been studied so far.

2 Results and Discussion

To investigate possible roles of nucleosome positioning in the recognition of *Paramecium* introns, we have determined the nucleosome occupancy profile along the *Paramecium* somatic genome by deep sequencing of DNA regions protected by nucleosomes against the microccocal nuclease (MNase) activity (Materials and Methods). We characterized the distribution, size and position of nucleosomes and linkers as well as nucleosome-free regions (NFR). This is the first study of nucleosome occupancy along an A/T-rich (70% AT) ciliate genome.

1.1 Nucleosome positioning along genes

Analysis of the nucleosome density profile showed that nucleosomes are highly organized along transcription units detected by deep sequencing of mRNA (Materials and Methods). As shown in figure 1A, the data reveal a specific gene promoter nucleosome pattern characterized by a nucleosome-depleted region (NFR) of typical size ~100bp upstream of the transcription start site (TSS) and, to a lesser extent, at the gene 3' end. These NFRs located at 5' and 3' gene ends serve as two excluding energy barriers at gene extremities. They define the intragenic chromatin structure and likely contribute to transcription initiation regulation. Nucleosomes follow an architecture that depends on the distance between the NFRs located at the 5' and 3' gene ends in a "crystal-like" array as observed in *Saccharomyces cerevisiae* [5]. However, we did not observe a specific enrichment of poly-T and poly-A tracts at gene 5' and 3' ends, respectively. The presence of such enrichment in another A/T-rich genome *Dictyostelium discoideum* has been suggested to contribute to the precise nucleosome positioning and transcription regulation in this organism [6].

1.2 Nucleosome positioning around introns

We analyzed nucleosome positioning around introns that were either predicted or determined by mRNA-seq. We observed that introns are frequently located within nucleosome linkers (Fig. 1B). In addition, nucleosome occupancy presents periodic distributions upstream (downstream) of 5’ (3’) intron borders, which does not result from the MNase digestion of AT-rich introns since MNase digestion of naked DNA did not generate these periodic distributions. Analyses of these distribution properties around various intron classes allow us to propose a model, in which coupling between RNA polymerase II elongation and nucleosome positioning at intron borders contributes to intron recognition and splicing. Our data strongly suggest that nucleosome positioning plays an important role in *Paramecium* intron evolution.

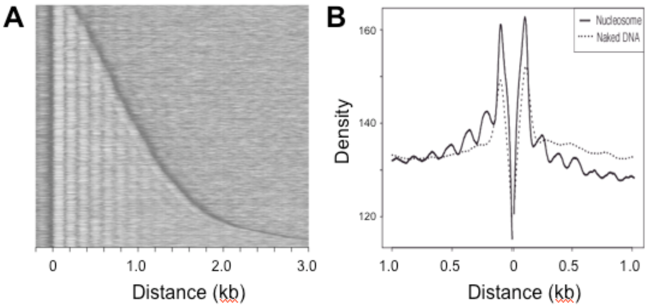


Figure 1. Nucleosome positioning along *Paramecium* genes (A) and around introns (B). (A) Heatmap of nucleosome occupancy profile of 10,031 genes with annotated transcription start (TSS) and termination (TTS) sites. Each gene is figured along a horizontal line and the grey level corresponds to the nucleosome occupancy density (dark grey corresponds to low nucleosome density). Genes were aligned by their TSS and were sorted in ascending length order. (B) Mean density profiles of nucleosome occupancy (solid line) and MNase digested naked DNA (dashed line) around 70,303 introns confirmed by mRNA-seq. Introns were aligned by their 5’ and 3’ borders, respectively.

3 Materials and Methods

Somatic nuclei from exponentially growing *P. tetraurelia* cells (strain 51) were isolated as described previously [7]. MNase was added to chromatin-containing or naked DNA (control). DNA associated with mono-nucleosomes or control DNA of the same size range was sequenced using the Illumina technique. RNA-seq libraries were generated according to manufacturer’s instructions using polyA+ RNAs extracted from cells at different stages of the life cycle as described previously [8].

Acknowledgements

This work has benefited from facilities and expertise of the IMAGIF sequencing platform (Centre de Recherche de Gif) and was supported by the ANR grant (ANR-10-BLAN-1603: GENOMAC).

References

[1] U. Braunschweig, S. Gueroussov, A.M. Plocik, B.R. Graveley and B.J. Blencowe, Dynamic Integration of Splicing within Gene Regulatory Pathways. *Cell*, 152: 1252-1269, 2013.

[2] S. Schwartz, E. Meshorer and G. Ast, Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, 16: 990-995, 2009.

[3] H. Tilgner, C. Nikolaou, S. Althammer, M. Sammeth, M. Beato, J. Valcarcel and R. Guigo, Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol*, 16: 996-1001, 2009.

[4] O. Jaillon, K. Bouhouche, J.F. Gout, J.M. Aury, B. Noel, B. Saudemont, M. Nowacki*, et al.*, Translational control of intron splicing in eukaryotes. *Nature*, 451: 359-362, 2008.

[5] C. Vaillant, L. Palmeira, G. Chevereau, B. Audit, Y. d'Aubenton-Carafa, C. Thermes and A. Arneodo, A novel strategy of transcription regulation by intragenic nucleosome ordering. *Genome Res*, 20: 59-67, 2010.

[6] G.S. Chang, A.A. Noegel, T.N. Mavrich, R. Muller, L. Tomsho, *et al.*, Unusual combinatorial involvement of poly-A/T tracts in organizing genes and chromatin in Dictyostelium. *Genome Res*, 22: 1098-1106, 2012.

[7] O. Arnaiz, N. Mathy, C. Baudry, *et al.*, The Paramecium germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS Genet*, 8: e1002984, 2012.

[8] O. Arnaiz, J.F. Gout, M. Betermier, K. Bouhouche, J. Cohen, L. Duret, *et al.*, Gene expression in a paleopolyploid: a transcriptome resource for the ciliate Paramecium tetraurelia. *BMC Genomics*, 11: 547, 2010.

Swarm: a Fast and Robust Clustering Method for Amplicon-based Studies

Frédéric MAHÉ[1], Torbjørn ROGNES[2,3,4], Colomban DE VARGAS[5] and Micah DUNTHORN[1]

[1] Department of Ecology, University of Kaiserslautern, Erwin-Schrödinger Str. 14, D-67663 Kaiserslautern, Germany
{mahe, dunthorn}@rhrk.uni-kl.de

[2] Department of Microbiology, Oslo University Hospital, Rikshospitalet Oslo University Hospital, P.O box 4950, Nydalen, 0424 Oslo, Norway

[3] Centre for Molecular Biology and Neuroscience (CMBN), Oslo University Hospital, Rikshospitalet Oslo University Hospital, P.O box 4950, Nydalen, 0424 Oslo, Norway

[4] Department of Informatics, University of Oslo, P.O box 1080, Blindern, 0316 Oslo, Norway
rognes@ifi.uio.no

[5] Adaptation et Diversité en Milieu Marin, UMR 7144 CNRS–UPMC, Université de Paris 6, Station Biologique de Roscoff, F-29680 Roscoff, France
vargas@sb-roscoff.fr

Keywords pairwise global alignment, molecular operational taxonomic units, biodiversity

1 Amplicon-based Studies & Traditional Clustering Methods

Identification of the total diversity of the smallest organisms—Archaea, Bacteria and microbial Eukaryotes—is a challenging task that necessitates the sequencing of one or several genomic markers (amplicons). With next generation sequencing technologies, many millions of amplicons can be obtained from a single environmental or clinical sample. To reduce this complexity for subsequent bioinformatics and statistical analyses, amplicons must be clustered into molecular operational taxonomic units (MOTUs), on the basis of their sequence similarities.

The clustering methods currently used have a similar approach: an amplicon is drawn from the amplicon pool and becomes the center of a new MOTU (centroid selection), this centroid is then compared to all other amplicons remaining in the pool. Amplicons for which the distance is within an arbitrary global clustering threshold to the centroid are moved from the pool to the MOTU, and the MOTU is closed. These steps are repeated as long as amplicons remain in the pool. This greedy approach suffers from two fundamental flaws: centroid selection induced input-order dependency, and arbitrary global clustering thresholds. The selection of an amplicon to become the center of a MOTU has important consequences on the final clustering results, but the relevance of that selection is not reevaluated throughout the clustering process. Clustering results are therefore dependent of the input-order of the amplicons. Finally, the use of a global clustering threshold does not fit well with the different rates of evolution observed among the branches of the tree of life. Swarm was developed to solve these issues.

2 Swarm: Fast and Robust Clustering

Swarm does not use a global similarity threshold to group amplicons, but iterates on a local clustering threshold calculated as the number of mismatches (substitutions or indels) between two amplicons once their optimal pairwise global alignment has been found. Our assumption is that clear cuts exist in the amplicon space; i.e. that amplicons do not form a continuum. If that condition is true, MOTUs can be allowed to grow iteratively until their natural limits are reached. Operating in that way removes the two main sources of variability of traditional clustering methods: Swarm outlines MOTU shapes without imposing one (no global threshold), and produce the same MOTUs regardless of the initial amplicon input-order.

To avoid comparing all pairs of amplicons, Swarm uses several computationally inexpensive filterings (sequence lengths, sequence compositions, triangular inequality) that greatly reduce the total computation time. To speed up the remaining pairwise comparisons, Swarm implements a multithreaded global pairwise alignment algorithm using the SSE4 instructions of modern CPUs (see Swipe for a similar implementation [5]).

The comparison of Swarm, Usearch [1], DNAclust [3] and CD-HIT-454 [2] results on a mock community (EMBL-EBI SRP003773) [4] showed that Swarm converges more rapidly on the expected number of MOTUs. Swarm results were not impacted by changes in the amplicon input-order, while Uclust and CD-HIT-454 results showed variability. Swarm also delivered better results in the regions of the amplicon-space where MOTUs are close from one another. In these dense regions, the global clustering threshold used by traditional methods can be too large. The first created MOTU subsumes its neighbors, leading to the formation of inaccurate MOTUs. When applied to large-scale environmental amplicon-based studies, such as BioMarKs (European coastal waters) and TARA OCEANS (global marine diversity), Swarm's high resolutive power yields fine grain MOTUs and reveals new patterns of molecular diversity.

3 Perspectives

Under certain circumstances, e.g. short and/or slowly evolving genomic markers, continuums of amplicons can appear. Typically, some amplicons form a bridge between two or several MOTUs, leading to the formation of a super-MOTU. Our tests on BioMarKs and TARA OCEANS showed that super-MOTUs are easy to detect, and concern only a small fraction of the MOTUs created by Swarm. We are currently working on a solution based on network analysis tools (articulation points, modularity) to automatically detect and break down super-MOTUs into their atomic components.

Swarm is available at `https://github.com/torognes/swarm`.

Acknowledgements

F.M. and M.D. were supported by the Deutsche Forschungsgemeinschaft (grant # DU1319/1-1). T.R. was supported by a Centre of Excellence grant from the Research Council of Norway to CMBN. C.d.V. was supported by the ANR POSEIDON. We are grateful to the CNRS-UPMC ABiMS bioinformatics platform (`http://abims.sb-roscoff.fr`) and to the Regional Computing Center at the University of Kaiserslautern for providing computational resources and support.

References

[1] R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26:2460-2461, 2010.

[2] L. Fu, B. Niu, Z. Zhu, S. Wu and W. Li. CD-HIT: Accelerated for Clustering the Next-generation Sequencing Data. *Bioinformatics*, 28:31503152, 2012.

[3] M. Ghodsi, B. Liu and M. Pop. DNACLUST: Accurate and Efficient Clustering of Phylogenetic Marker Genes. *BMC Bioinformatics*, 12:271, 2011.

[4] C. Quince, A. Lanzen, R. J. Davenport and P. J. Turnbaugh. Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics*, 12:38, 2011.

[5] T. Rognes. Faster Smith-Waterman Database Searches with Inter-sequence SIMD Parallelisation. *BMC Bioinformatics*, 12:221, 2011.

Computational modeling of T helper cell differentiation

Wassim ABOU-JAOUDE[1], Maximilien GRANDCLAUDON[2], Vassili SOUMELIS[2] and Denis THIEFFRY[1]

[1] Institut de Biologie de l'Ecole Normale Supérieure, UMR ENS, CNRS 8197, INSERM 1024,
46 rue d'Ulm, F-75230, Paris, Cedex 05, France
wassim@biologie.ens.fr, thieffry@ens.fr

[2] Institut Curie, Laboratoire d'Immunologie Clinique, INSERM U653,
26 rue d'Ulm, F-75248, Paris, Cedex 05, France
{maximilien.grandclaudon, vassili.soumelis}@curie.net

Keywords logical modeling, systems immunology, T-helper cell differentiation and plasticity.

1 Introduction

T helper (CD4+) lymphocytes play a key role in the regulation of immunity. Potentially faced with a large diversity of microbial pathogens, antigen-inexperienced naive CD4+ T cells orchestrate immune responses after differentiating into T helper (Th) cell populations that secrete distinct sets of cytokines. This differentiation process requires the integration of multiple signals, produced by the actors of the immune response, triggering specific surface receptors, including the T cell receptor, co-stimulatory molecules, and cytokine receptors. Diverse combinations of these signals lead to the differentiation of naïve T cells into diverse Th subsets, among which Th1, Th2, Treg and Th17 subtypes. This diversity of response has evolved in order to best control the diversity of pathogens encountered. For example, parasites and extracellular bacteria induce the differentiation of naive T cells into Th2 cells leading to microbial clearance. In this manner, Th cells can tailor their responses to the nature of the threat.

We have recently proposed a comprehensive dynamical model to account for the differentiation of Th cells to Th1, Th2, Treg and Th17 subtypes [1]. Relying on a sophisticated logical modelling framework, model simulations enabled to assess the effects of heterogeneous environments on Th cell differentiation. This led to identify stable states corresponding to canonical Th1, Th2, Th17 and Treg subtypes, but also to hybrid cell types co-expressing combinations of Th1, Th2, Treg and Th17 markers in an environment-dependent manner. Altogether, these computational studies highlight the diversity and plasticity of Th lymphocytes.

2 Aim and methods

Based on [1], our goal is to propose an extended and refined logical model of Th differentiation and plasticity. This extended model will integrate novel data produced on the response of Th cells to various cytokine environments. Novel computational methods (for model reduction, state transition graph compression and regulatory circuit analysis implemented in the logical modelling software *GINsim*) will be used to cope with the large number of components composing the model (of the order of 100) and generate meaningful results, emphasizing the attractors, corresponding to Th subsets, and the most important transitions underlying commitment [1, 2]. Simulations and analyses of the dynamical model will contribute to decipher the mechanism of Th differentiation and suggest novel experiments to design in order to test these predictions.

3 First results

We have started to extend and refine Naldi's model to integrate recently published data on the response of Th cells to several cytokine combinations in terms of cytokine expression patterns [3, 4]. The following cytokines have been incorporated into the model and wired to the intracellular signaling components: (i) IL-1β (notably involved in human Th17 differentiation) and interferon (IFN)-α as input cytokines; (ii) IL-5, IL-13 (which are part of the cytokine signature of Th2 cells), IL-22 and IL-6 as secreted output cytokines. The analysis of the resulting logical model is in progress. The refined model already reproduced additional important cytokine patterns observed in [3, 4], such as the differential regulation of IL-22 and IL-6 production in Th1 conditions, and the differential expression of IL-21 in Th1 and Th2 conditions.

4 Prospects

Simulations will be performed in order to account for the other cytokine expression patterns features observed in [3, 4]. Additional input and output cytokines will be progressively included into the model, along with important signaling elements underlying input-output cytokines relationships.

Once the model made coherent with existent data, systematic simulations will be performed to generate meaningful predictions. The effects of different combinations of cytokines in different temporal orders will be simulated to predict the impact of context and timing variations on cellular decision. To gain insight into the temporal response of the system, the dynamical behaviour of the model will be investigated. In particular, transient activation and repression of cytokines and transcription factors will be assessed. Reprogramming from one differentiated Th subtype into others in response to various environmental input conditions will also be investigated to assess the stability of the Th lineages.

Simulations of single or multiple genetic perturbations (knock-downs, ectopic gene expressions) will be systematically performed to characterize the roles of critical components involved in Th commitment and plasticity. Tentatively these components will then be targeted experimentally to validate their importance in the Th differentiation process, and the results of these experimental studies should in turn help us to improve the fitness and predictive power of the model.

5 Conclusion

Ultimately, this work should enlighten the basic mechanisms regulating T cell differentiation and plasticity by multiple stimuli reflecting the complexity of inflammatory environments. Furthermore, it should improve our ability to design therapeutic strategies targeting specific microbial agents.

Acknowledgements

This work was supported by the Labex MemoLife.

References

[1] A. Naldi, J. Carneiro, C. Chaouiya and D. Thieffry, Diversity and plasticity of Th cell types predicted from regulatory network modelling. *PLoS Comput. Biol.*, 6: e1000912, 2010.

[2] C. Chaouiya, A. Naldi and D. Thieffry, Logical modelling of gene regulatory networks with GINsim. *Meth. Mol. Biol.,* 804: 463-79, 2012.

[3] E. Volpe, N. Servant, R. Zollinger, S.I. Bogiatzi, P. Hupé, E. Barillot and V. Soumelis, A critical function for transforming growth factor-beta, interleukin 23 and proinflammatory cytokines in driving and modulating human T(H)-17 responses. *Nat. Immunol.,* 9: 650-7, 2008.

[4] E. Volpe, M. Touzot, N. Servant, M.A. Marloie-Provost, P. Hupé, E. Barillot and V. Soumelis, Multiparametric analysis of cytokine-driven human Th17 differentiation reveals a differential regulation of IL-17 and IL-22 production. *Blood,* 114: 3610-4, 2009.

The Bcl-2 family database (BCL2DB): a mixed bag for a family of four

Valentine RECH DE LAVAL[1], Gilbert DELEAGE[1], Abdel AOUACHERIA[2] and Christophe COMBET[1]

[1] Bases Moléculaires et Structurales des Systèmes Infectieux, UMR5086 CNRS/UCBL,
7 passage du Vercors, 69 367, Lyon, Cedex 07, France
christophe.combet@ibcp.fr

[2] Laboratoire de Biologie Moléculaire de la Cellule, UMR5239 CNRS/ENS Lyon/UCBL/HCL,
46, allée d'Italie, 69364, Lyon, Cedex 07, France
abdel.aouacheria@ens-lyon.fr

Keywords Apoptosis, Bcl-2 family, sequence motifs, remote homology, database.

1. Introduction

Bcl-2 family members are essential regulators of apoptosis (cell death by suicide) and probably of other cellular processes as well [1]. This family is formed by a group of proteins homologous (i.e., evolutionary linked) to the founding member, Bcl-2, and by a collection of evolutionarily and structurally unrelated proteins characterized by the presence of a single region of sequence similarity with Bcl -2, termed the BH3 motif. Bcl-2 homologous proteins share a similar all-α helical bundle fold (the 'Bcl-2 domain'), have up to four different BH motifs (BH3 plus BH1, BH2 and BH4), and can be either anti-apoptotic (The Good, e.g., Bcl-2 and Bcl-xL,) or pro-apoptotic (The Bad, e.g., Bax, Bak and Bid), while all of the BH3-only proteins (The Ugly) are pro-apoptotic. A variety of viral proteins (the remaining members of the family, The Stranger) have been found to be structurally similar to Bcl-2 with or without obvious sequence similarity, suggesting that BH motifs are only part of the picture [2].

Since the discovery of the *bcl-2* gene twenty-nine years ago, intense research in various disciplines has exponentially increased the quantity of data available about Bcl-2 family proteins. It is therefore of considerable interest to use bioinformatic tools to (i) understand the various subgroups forming the Bcl-2 family and their implication in diseases; (ii) bring all the available information together in a specialized database. We recently proposed a novel classification scheme for Bcl-2 family proteins, based on phylogenetic information and computational analysis of sequence data [3,4]. Here, we describe the Bcl-2 Family Database (http://bcl2db.ibcp.fr/), a database-backed web resource on the Bcl-2 family for which a prototype was previously generated [5].

2. Results

Based on our new family classification, we built an automated workflow to feed the Bcl-2 family database (BCL2DB), a collection of computer-annotated Bcl-2 family sequences. The workflow starts by a profile search against the UniProt Knowledgebase (UniProtKB, release 2012_11) [6] using a suite of in-house programs (including the 'FindBcl2' and 'FindBH3' algorithms) implemented in Java. This computational pipeline utilizing information derived from specific profiles of proteins [7] was able to identify close and distant homologs of Bcl-2 (including viral members) as well as the known repertoire of BH3-only proteins. A total number of 2210 UniProtKB matches were found, yielding 1324 unique BCL2DB entries. For the detected sequences, cross-reference queries and BLAST searches retrieved corresponding nucleotide sequence entries from Ensembl [8], Ensembl Genomes [9] or European Nucleotide Archive (ENA) [10]. Unwanted annotations were deleted from the entries in order to provide BCL2DB entry templates. Then, the templates were filled with supplemental annotations (e.g. subfamily, paralogy group and BH motif composition) using the 'AnnotateBCL2' algorithm and loaded into the PostgreSQL relational database management system (RDBMS). The annotation process automatically affiliates each identified protein to its closest homology group based on a specific, curated gathering threshold cut-off (different for each profile). Beyond the threshold, the entry is considered as 'unclassified'. Moreover, homemade BH motif profiles (PSSM) were developed for use in computational annotation of BCL-2 family sequences. The relational database can be queried via a web interface and the query results can be further analyzed with generic (alignment, homology search) and specialized (annotation) integrated analysis tools. The generic analysis tools (e.g. BLAST or Clustal W) are available through the NPS@ web server [11], which is an integrated computational work bench. Multiple sequence alignments (calculated using the program MUSCLE [12]) and other sequence- or structure-derived data sorted by family member were pre-computed and stored on-site for access through a web interface at http://bcl2db.ibcp.fr. The sequence alignments (displayed in Clustal W format) can be interactively edited with the 'EditAlignment' applet developed by our team [13]. Users can also export output datasets in the form of Pearson/Fasta sequences, accession number lists and entry flat files for further analysis. BCL2DB is updated on a monthly basis from the UniProtKB, Ensembl, Ensembl Genomes, EMBL, InterPro and PDB databases.

3. Conclusion and perspectives

The large and heterogeneous mixture of proteins forming the extended Bcl-2 family poses major challenges to researchers trying to investigate structure-function relationships, molecular evolution and other bioinformatics topics, with the risk of hampering further scientific and biomedical progress in the field. The Bcl-2 Family Database represents an effort toward providing a global picture of Bcl-2 family proteins and making tools available for their analysis. In the future, this resource will be enriched with novel data types: gene co-expression, genetic variation, evolutionary conservation and protein-protein interactions, and new tools will be provided (for instance homology modeling of protein 3D structures and visualization of annotated phylogenetic trees).

Acknowledgements

V.R.L. is the recipient of a doctoral fellowship from La Ligue Contre le Cancer (Comité de Saône-et-Loire). BCL2DB is developed on the Pôle Rhône-Alpes de BioInformatique platform funded by the Groupement d'Intérêt Scientifique Infrastructures en Biologie Sante et Agronomie (GIS IBiSA-AO 2009). We are grateful to Région Rhône-Alpes for a grant (CIBLE program).

References

[1] J.M. Hardwick and L. Soane, Multiple functions of BCL-2 family proteins. *Cold Spring Harb Perspect Biol.,* 5, 2013.

[2] J.M. Hardwick and R.J. Youle, SnapShot: BCL-2 proteins. *Cell.*, 138:404, 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.,* 22:2395-2416, 2005.

[4] A. Aouacheria, V. Rech de Laval, C. Combet and J.M. Hardwick, Evolution of Bcl-2 homology motifs: homology versus homoplasy. *Trends Cell Biol.*, 23:103-111, 2013.

[5] S.V. Blaineau, A. Aouacheria. BCL2DB: moving 'helix-bundled' BCL-2 family members to their database. *Apoptosis*, 14(7):923-5, 2009.

[6] UniProt Consortium, Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, 41:D43-D47, 2013.

[7] A. Krogh, M. Brown, I.S. Mian, K. Sjölander and D. Haussler, Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol.*, 235:1501-1531, 1994.

[8] P. Flicek, I. Ahmed, M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, *et al.*, Ensembl 2013. *Nucleic Acids Res.*, 41:D48-D55, 2013.

[9] P.J. Kersey, D.M. Staines, D. Lawson, E. Kulesha, P. Derwent, J.C. Humphrey, D.S. Hughes, S. Keenan, A. Kerhornou, G. Koscielny, N. Langridge, M.D. McDowall, K. Megy, U. Maheswari, M. Nuhn, M. Paulini, H. Pedro, I. Toneva, D. Wilson, A. Yates and E. Birney, Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, 40:D91-D97, 2012.

[10] V. Zalunin, Facing growth in the European Nucleotide Archive. *Nucleic Acids Res.*, 41:D30-D35, 2013.

[11] C. Combet, C. Blanchet, C. Geourjon and G. Deléage, NPS@: network protein sequence analysis. *Trends Biochem Sci.,* 25:147-150, 2000.

[12] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32:1792-1797, 2004.

[13] C. Combet, N. Garnier, C. Charavay, D. Grando, D. Crisan, J. Lopez, A. Dehne-Garcia, C. Geourjon, E. Bettler, C. Hulo, *et al.*, euHCVdb: the European hepatitis C virus database. *Nucleic Acids Res.,* 35:D363-D366, 2007.

Deciphering genome-wide cis-regulations with RSAT

Morgane THOMAS-CHOLLIER[1], Matthieu DEFRANCE[2], Alejandra MEDINA-RIVERA[3], Olivier SAND[4], Pierre VINCENS[1], Carl HERRMAN[5], Denis THIEFFRY[1] and Jacques VAN HELDEN[5]

[1] Computational systems biology, Institute of Biology of ENS (IBENS), Paris, France
mthomas@biologie.ens.fr , thieffry@ens.fr , Pierre.Vincens@ens.fr

[2] Laboratory of Cancer Epigenetics, Faculty of Medicine, Université Libre de Bruxelles, Belgium
defrance@bigre.ulb.ac.be

[3] SickKids Research Institute, 101 College St. East Tower | Suite 15-306, Toronto, Ontario, Canada
alejandra.medina@sickkids.ca

[4] Génomique et maladies métaboliques, CNRS-UMR8199, Institut de Biologie de Lille, France
olivier.sand@good.ibl.fr

[5] Technological Advances for Genomics and Clinics (TAGC), INSERM U928 & Aix-Marseilles University, France
jacques.VAN-HELDEN@univ-amu.fr , carl.herrmann@univ-amu.fr

Keywords cis-regulatory elements, motif discovery, ChIP-seq

The regulatory sequence analysis tools (RSAT, http://rsat.fr and mirrors) is a software suite that integrates a large collection of modular tools for the detection of cis-regulatory elements in genomic sequences [1-3]. The web site has been running without interruption since 1998, and the suite has been continuously developed to accommodate novel types of data and experimental approaches over the years [2-6]. The suite includes programs for sequence retrieval, motif discovery, phylogenetic footprint detection, sequence scanning with regular expressions or position-specific scoring matrices, motif quality assessment and comparison, visualization and conversion utilities, along with a series of tools for random model generation and statistical evaluation. Genomes are regularly updated from various genome repositories (NCBI, Ensembl, UCSC browser) and the website currently supports 2517 genomes (March 2013).

RSAT enables genome-wide analysis of cis-regulatory elements with different types of input data: (i) groups of co-expressed genes produced by transcriptomic experiments, (ii) phylogenetically conserved regions and (iii) high-throughput binding data such as ChIP-seq.

Several efficient and complementary motif discovery algorithms can predict transcription factor binding motifs from groups of co-expressed genes. Although these methods yields good results in yeast and bacteria genomes, they are not suitable for vertebrates, due to the larger size and heterogeneity of non coding genomic sequences. The same algorithms nevertheless proved very efficient to analyse high-throughput transcription factor binding data, where the signal to noise ratio is higher. The workflow *peak-motifs* [6] was therefore developed to process large collections of peak sequences obtained from ChIP-seq or related technologies, to predict transcription factor binding motifs, match them against motif databases and predict their binding sites. Current developments aim at extending *peak-motifs* to support other high-throughput data, including chromatin marks, DNaseI, or ChIP-exo profiles.

Sequence conservation across species is often used to reduce the search space for cis-regulatory elements. In yeast and bacteria, RSAT supports phylogenetic footprinting to detect conserved motifs in promoters of orthologous genes [7, 8]. We are currently completing the pattern-matching approaches to take into account phylogenetic conservation among larger genomes (e.g. from UCSC browser) and thereby focus on the predicted binding sites that are the most likely functional.

In addition to motif discovery, the suite already provides various tools to analyse transcription factor binding motifs represented as matrices, including motif comparisons [3] and evaluation of matrix quality [5]. We are currently developing a motif clustering algorithm to ease the analysis of overlapping motifs (newly discovered or reported in databases) and assess potential motif diversity for a given transcription factor, in the context of motifs for cofactors.

RSAT web server offers an intuitive interface, where each program can be accessed either separately or connected to the other tools. In addition, many tools are available as SOAP/WSDL web services, enabling their integration in programmatic workflows. Programs are documented with manual pages, while 'demo' buttons propose typical test cases. In addition, web tutorials and a series of published protocols help the users to master the different functionalities of RSAT [9-12], providing step-by-step guidelines about alternative options s, as well as regarding the interpretation of results.

References

[1] van Helden J. Regulatory sequence analysis tools. *Nucleic Acids Research* 31: 3593-6, 2003.

[2] Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E., Brohee S & van Helden J. RSAT: regulatory sequence analysis tools. Nucleic Acids Res*earch* 36: W119-27, 2008.

[3] Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J. RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Research* 39: W86-91, 2011.

[4] Defrance M, & van Helden J. Info-gibbs: a motif discovery algorithm that directly optimizes information content during sampling. *Bioinformatics* 25: 2715-22, 2009.

[5] Medina-Rivera A, Abreu-Goodger C, Salgado-Osorio H, Collado-Vides J & van Helden J. Empirical and theoretical evaluation of transcription factor binding motifs. *Nucleic Acids Research* 39: 808–24, 2011.

[6] Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J.RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research* 40: e31, 2012.

[7] Janky R & van Helden J. Evaluation of phylogenetic footprint discovery for the prediction of bacterial cis-regulatory elements. *BMC Bioinformatics* 9: 37, 2008.

[8] Brohee S, Janky R, Abdel-Sater F, Vanderstocken G, Andre B & van Helden J. Unraveling networks of co-regulated genes on the sole basis of genome sequences. *Nucleic Acids Research* 39: 6340-58, 2011.

[9] Turatsinze JV, Thomas-Chollier M, Defrance M & van Helden J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols* 3 : 1578-88, 2008.

[10] Defrance M, Janky R, Sand O.& van Helden J. Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nature Protocols* 3 : 1589-603, 2008.

[11] Sand O, Thomas-Chollier M, Vervisch E & van Helden J. Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services-an example with ChIP-chip data. *Nature Protocols* 3 : 1604-15, 2008.

[12] Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D & van Helden J. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature Protocols* 7: 1551-68, 2012.

Sélection positive chez les gènes récemment dupliqués dans les génomes de plantes

Looking for positive selection in recently duplicated genes in plant genomes

Iris Fischer[2], Jacques Dainat[1], Jean-François Dufayard[3], Vincent Ranwez[1] and Nathalie Chantret[2]

[1] AGAP, UMR1334 SupAgro, 2 Place P. Viala, 34060, Montpellier, Cedex 1, France
{j.dainat, vincent.ranwez}@supagro.inra.fr

[2] AGAP, UMR1334 INRA, 2 Place P. Viala, 34060, Montpellier, Cedex 1, France
{iris.fischer, nathalie.chantret}@supagro.inra.fr

[3] AGAP, UMR1334 CIRAD-BIOS, Avenue Agropolis, 34398, Montpellier, Cedex 5 France
jeanfrancois.dufayard@gmail.com

Mots-clés Duplication, génomes, sélection positive, paralogue, orthologue.

Les duplications de gènes sont des évènements importants dans l'évolution des génomes, et sont particulièrement fréquents dans les différentes lignées de plantes. Dans la majorité des cas après un évènement de duplication, dû à la redondance des fonctions, un des deux gènes adopte une évolution neutre et disparait par pseudogénisation. Dans certains cas la nouvelle copie se fixe (Proost, Pattyn, Gerats, & Van de Peer, 2011) : elle peut garder sa fonction d'origine (et augmenter de ce fait la transcription) ; elle peut se sub-fontionnaliser (Les deux copies se partagent la fonction initiale) ; elle peut se néo-fonctionnaliser (acquérir une nouvelle fonction ou spécialiser son profil d'expression). Il a été suggéré qu'une large part de la diversité chez les plantes est causée par la duplication de gènes et la spécialisation adaptative des copies paralogues (Flagel & Wendel, 2009). Cette hypothèse d'évolution adaptative des paralogues peut se caractériser par l'action de la sélection positive.

Dans notre travail nous avons étudié la sélection positive au sein de gènes récemment dupliqués chez les plantes. En tirant l'avantage des derniers progrès dans le séquençage de génome complet de plantes, nous avons choisi de porter notre étude sur 10 génomes d'angiospermes. Notre travail se concentre sur l'analyse de groupes de paralogues contenant au moins 6 copies (**Ultraparalogues**). Pour détecter les ultraparalogues au sein des 10 espèces investiguées, nous avons utilisé l'étude phylogénétique des familles de gènes préalablement clusterisées à l'aide de TribeMCL (Enright, Van Dongen, & Ouzounis, 2002), disponible dans la base de données GreenPhylDB. L'analyse de ces données nous a permis de mettre en évidence 2781 groupes d'Ultraparalogues lignée spécifique. Afin d'étudier les pressions de sélection au sein de ces groupes, les séquences présentes dans chacun de ces groupes ont été alignées. Les alignements ont été effectués en utilisant la combinaison de l'algorithme d'alignement PRANK (Löytynoja & Goldman, 2005) qui prend comme guide la topologie d'un arbre d'espèces, et de l'outil GUIDANCE (Penn, Privman, Landan, Graur, & Pupko, 2010) qui permet de quantifier la robustesse des colonnes et des séquences de l'alignement. GUIDANCE nous permet ainsi de nettoyer le signal des alignements. Il a été montré que la combinaison de ces deux outils permet d'obtenir les alignements les plus fiables possibles afin d'analyser par la suite la sélection positive en utilisant les modèles d'évolution des codons implémentés dans PAML (Fletcher & Yang, 2010; Jordan & Goldman, 2012). L'analyse des pressions de sélection se base sur le calcul de ω qui est le ratio du nombre de substitutions non-synonymes par sites non-synonymes (dN) sur le ratio du nombre de substitutions synonymes par sites synonymes (dS). Un ratio ω supérieur à 1 indique la présence de sélection positive.

Les alignements obtenus ont été étudiés dans un premier temps pour détecter de la sélection positive par site à l'aide du programme CODEML (Yang, 2007). Dans un deuxième temps nous avons effectué une recherche de sélection positive par branche grâce la mise au point d'un workflow basé sur l'outil MapNH (Dutheil et al., 2012) qui permet un calcul robuste du ω avec un temps de calcul fortement optimisé par rapport à l'utilisation de CODEML.

L'étude des pressions de sélection a également été effectuée sur 2407 groupes d'orthologues stricts témoins, provenant des mêmes arbres phylogénétiques que les groupes d'ultraparalogues. L'étude des pressions de sélection au sein des groupes d'orthologues et d'ultraparalogues nous ont permis de mettre en évidence des différences significatives entre ces deux groupes.

Seulement 0,4% des groupes d'orthologues ont des sites sous sélections positives, contre 5,8% pour les groupes d'ultraparalogues. La proportion de branches ayant un ω>1 est de 0,3 % pour les orthologues et de 12% pour les paralogues.

Nous avons une forte corrélation entre les alignements où nous trouvons des sites sous sélection positive et ceux où nous trouvons des branches sous sélection positive. En effet, lorsque l'on considère uniquement les branches des alignements pour lesquels de la selection positive par site a été détectée, nous augmentons le pourcentage de branches sous sélection positive de 12% à 23%.

L'analyse de la moyenne du ratio ω par branche nous montre une différence significative entre les groupes d'orthologues et d'ultraparalogues (respectivement 0,28 et 0,61). L'étude de la distribution des ratios ω pour ces deux groupes nous montre un relâchement de la sélection purificatrice dans les groupes d'ultraparalogues.

References

[1] Dutheil, J. Y., Galtier1, N., Jonathan, R., Douzery, E. J. P., Ranwez, V., & Boussau, B. (2012). Efficient Selection of Branch-Specific Models of Sequence Evolution. *Molecular Biology and Evolution*. Retrieved January 8, 2013.

[2] Enright, a J., Van Dongen, S., & Ouzounis, C. a. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, *30*(7), 1575–84.

[3] Flagel, L. E., & Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *The New phytologist*, *183*(3), 557–64.

[4] Fletcher, W., & Yang, Z. (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Molecular biology and evolution*, *27*(10), 2257–67.

[5] Jordan, G., & Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular biology and evolution*, *29*(4), 1125–39.

[6] Löytynoja, A., & Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(30), 10557–62.

[7] Penn, O., Privman, E., Landan, G., Graur, D., & Pupko, T. (2010). An alignment confidence score capturing robustness to guide tree uncertainty. *Molecular biology and evolution*, *27*(8), 1759–67.

[8] Proost, S., Pattyn, P., Gerats, T., & Van de Peer, Y. (2011). Journey through the past: 150 million years of plant genome evolution. *The Plant journal : for cell and molecular biology*, *66*(1), 58–65.

[9] Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–91.

Mobile Genetic Elements : Major Providers of Genetic Public Goods in the Human Gut

Cedric Bicep[1], Eric Bapteste[1] and Philippe Lopez[1]

[1] LABORATOIRE SAE, UMR 7138 CNRS, UPMC, 75005 Paris, France
cbicep@snv.jussieu.fr

Keywords mobile genetic element, metagenomics, microbiome, network.

1 Introduction

In the literature human body has been described as "Gestalt" entity that consists in both prokaryotic and eukaryotic cells [1]. It has been established that our human body harbours 10 times more bacteria than human cells. To describe the role of this community, the hologenome theory of evolution has been proposed by Zilber-Rosenberg & Rosenberg [2]. The hologenome is defined as the sum of the genetic information of the host and its microbiota. Our eukaryotic genome is fixed for life and lack plasticity in comparison to our harbored prokaryotic genomes which are more flexible. It is commonly admitted that prokaryotes are able to acquire new gene from horizontal gene transfer. Gene transfer is sometime mediated by mobile genetic element (MGE) such as plasmids, viruses and integrons. MGE infecting microbes living in our gut are known to play role in the spread of antibiotic resistance gene for example [3]. Those resistance genes can be considered as public goods since they can be acquired by any bacteria.

The extent of potentially mobilizable gene families of the human gut microbiome, and the nature of the selective pressure on them still remain poorly known. Our goal is to retrieve and analyze potentially mobile gene families of the human gut microbiota using sequence networks.

2 The method

After having predicted gene from contigs of the human gut microbiome dataset with MetaGeneAnnotator [4], we BLASTed all gene sequence against each other. This sequence analysis has been used to create sequences similarity networks. Sequence similarity networks are graphs where nodes represent individual gene sequence [5-10]. Those nodes are connected by edges when they display more than a certain similarity threshold (BLAST e-value < 1e-5, > 20% identity). The resulting network show many disconnected subnetworks or "connected components" which represent "operational" gene families. Those gene families present various sizes and topologies. Since graphs are mathematical objects, the topology of these connected components can be described with centralities measures, such as a clustering coeficient and graph properties, such as average clustering coefficient and variance to mean ratio (VMR) of the similarity between sequences belonging to the same gene family.

We used sequence similarity networks to determine potential gene families carried by mobile genetic elements (MGE). We applied this method to two metagenomic datasets of american and japanese gut microbiomes.

3 First experimental results

Our analysis reveals that these two microbiomes do not exhibit the same ratio of potentially mobile gene families. American gut microbiome are depleted in potentially mobile gene families compared to Japanese gut microbiome. We analyzed the functional distribution of potentially mobile and “non-mobile” gene families of the human gut. We found that gene families potentially carried by MGE are enriched in gene involved in metabolism and information storage processing. Our results suggest that MGE play the role of "public goods providers" to the bacterial community of the human gut, by providing them with essential genes and genetic diversity to survive in the human gut.

References

[1] Brian V Jones, The human gut mobile metagenome a metazoan perspective Gut Microbes. 2010 Nov-Dec; 1(6): 415–431. doi: 10.4161/gmic.1.6.14087 PMCID: PMC3056110.

[2] Zilber-Rosenberg I, Rosenberg E. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. FEMS Microbiol Rev. 2008 Aug; 32(5):723-35.

[3] Shoemaker NB, Vlamakis H, Hayes K, Salyers AA. Evidence for extensive resistance gene transfer among Bacteroides spp. and among Bacteroides and other genera in the human colon. Appl Environ Microbiol. 2001;67:561–568.

[4] Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res. 2008 Dec;15(6):387-96. doi: 10.1093/dnares/dsn027. Epub 2008 Oct 21.

[5] Bittner L, Halary S, Payri C et al. Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. Biol Direct 2010; 5: 47.

[6] Dagan T, Martin W. Getting a better picture of microbial evolution en route to a network of genomes. Phil Trans R Soc Lond B Biol Sci 2009; 364: 2187–2196.

[7] Fondi M, Fani R. The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks. Environ Microbiol 2010; 12: 3228–3242.

[8] Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. Network analyses structure genetic diversity in independent genetic worlds. Proc Natl Acad Sci USA 2010; 107: 127–132.

[9] Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Reticulate representation of evolutionary and functional relationships between phage genomes. Mol Biol Evol 2008; 25: 762–777.

[10] Beauregard-Racine J, Bicep C, Schliep K, Lopez P, Lapointe FJ, Bapteste E. Of woods and webs: possible alternatives to the tree of life for studying genomic fluidity in E. coli. Biol Direct 2011; 6: 39.

Fiona: a parallel and automatic strategy for read error correction

Marcel H. Schulz[1,4], David Weese[2,4], Manuel Holtgrewe[2,4], Viktoria Dimitrova[3], Sijia Niu[3], Knut Reinert[2], and Hugues Richard[3,4]

[1] Ray and Stephanie Lane Center for Computational Biology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA - maschulz@andrew.cmu.edu

[2] Department of Mathematics and Computer Science, Freie Universität Berlin, Takustr. 9, 14195 Berlin, Germany

[3] UPMC, Lab. de Gén. des Microorganismes UMR7238 CNRS, Génomique Analytique, F-75006 Paris, France. hugues.richard@upmc.fr

[4] These authors contributed equally to this work.

Abstract *We present Fiona, a new stand-alone read error correction method which provides a new statistical approach for sequencing error detection, optimal error correction and estimates its parameters automatically. Fiona is able to correct mismatch, insertion, and deletion errors and can be applied to any sequencing technology. When comparing its performance to state of the art methods, Fiona shows a constantly high correction accuracy over a broad range of datasets and shows its superiority in the presence of short indel errors, which are prevalent in 454 and Ion Torrent sequencing data. Fiona was implemented in the SeqAn open source C++ library for sequence analysis and is publicly available for download at:* *http://www.seqan.de/projects/fiona*.

Keywords Next Generation Sequencing, Error correction, Genome Assembly, Suffix Arrays.

Next generation DNA sequencing (NGS) has revolutionized genomics. NGS technologies produce millions of sequencing reads of a few hundred bases in length. Its more common use is for genome sequencing, which has important applications like the *de novo* genome assembly of previously unknown genomes or the re-sequencing of already assembled genomes and the analysis of sequence polymorphisms. Due to the vast amount of applications of genome sequencing, the correction of mismatch errors as well as insertion or deletion errors (indels) introduced by the sequencing machine has recently attracted attention. Previous studies showed that error correction before analysis can improve *de novo* genome assembly performance. However, current error correction approaches suffer a number of limitations as highlighted in a recent review about the subject [9]. *(i)* Most methods cannot correct indel errors because they are tailored to correct only mismatch errors and are therefore only applicable to Illumina reads. *(ii)* Most approaches need to be parameterized depending on the dataset, otherwise their performance is suboptimal. This either requires in-depth knowledge by the user or parameter optimization using downstream analysis which often leads to longer running times in practice. *(iii)* Because the throughput of NGS technologies is growing steadily many approaches not applicable to larger datasets because of running time or memory limitations. These caveats make it hard for users to choose the optimal tool for their dataset and NGS technology.

Here we introduce a new approach to read error correction, called Fiona, which addresses all the above mentioned limitations. Fiona provides an accurate and highly parallelized method for correction, with the ability to correct indel errors, while it automatically adjusts its parameters. The algorithm uses a suffix tree to detect and correct mismatch and indel errors following [8,6]. In the implementation, the suffix tree traversal is emulated using solely a partial suffix array that can be easily be constructed and traversed in parallel. Instead of treating discovered errors independently, it collects them and solves a new formulation for optimal error correction inspired by the MSA based correction methods. Further, it uses new statistical methods to improve error detection at different k values and automatically estimates its parameters inspired by [3]. Fiona was implemented in the SeqAn [1] open source C++ library for sequence analysis. The main steps of the Fiona algorithm are summarized in Fig. 1.

We extensively compared the performance of Fiona to state-of-the-art tools, namely Coral 1.4 [7] , ECHO 1.12 [4], HiTEC 1.0.2[3], Quake 0.3.4.2 [5], and the error correction module of Allpaths-LG 44994 [2] and on multiple

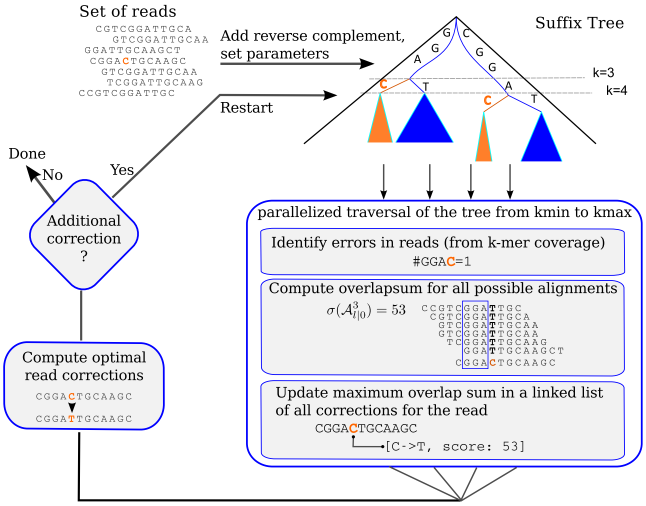


Figure 1. The Fiona strategy illustrated on a toy example. A suffix tree (in fact a partial suffix array) is built from the set of reads and their reverse complement and traversed to detect and correct errors. Potential errors in the reads are identified as nodes in the tree according to their coverage for various suffix lengths (for instance GGAC is covered by only one read). The correction with highest overlap score is chosen to correct the read at that position. As the tree is traversed in parallel, all possible corrections on a read are recorded in a linked list, which reports the positions of corrections as well as the current maximal value of the score. After traversal, the reads are updated with the correction of maximum score. Once all reads have been corrected, the algorithm decides if an additional round of correction should be performed

datasets, produced with various sequencing technology, Illumina (10 datasets), 454 (3 datasets) and Ion Torrent (1 dataset). We selected organisms with genomes of varying length and complexity, from short genomes (*E. coli*, *H. syringae*) to longer and more complex ones (*D. melanogaster*, *C. elegans*).

Fiona showed performance competitive with state-of-the-art methods on Illumina data and was robust against cases where the sequence coverage is lower. Fiona revealed a clear superiority to all methods when evaluated on 454 and Ion Torrent datasets. Our evaluation also showed that Allpaths-LG is a very good alternative considering running time and memory usage on high-coverage Illumina datasets. We believe that users will improve their downstream analysis by using Fiona in their pipelines and made Fiona publicly available under an open source license at: http://www.seqan.de/projects/fiona.

References

[1] Andreas Döring, David Weese, Tobias Rausch, and Knut Reinert. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9:11, 2008.

[2] *Gnerre et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, 108(4):1513–1518, 2010.

[3] Lucian Ilie, Farideh Fazayeli, and Silvana Ilie. HiTEC: accurate error correction in high-throughput sequencing data. *Bioinformatics*, 27(3):295–302, 2011.

[4] W.C Kao, A.H Chan, and Y.S Song. ECHO: A reference-free short-read error correction algorithm. *Genome Research*, 21(7):1181, 2011.

[5] David R Kelley, Michael C Schatz, and Steven L Salzberg. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol*, 11(11):R116, Nov 2010.

[6] Leena Salmela. Correction of sequencing errors in a mixed set of reads. *Bioinformatics*, 26(10):1284–1290, May 2010.

[7] Leena Salmela and Jan Schröder. Correcting errors in short reads by multiple alignments. *Bioinformatics*, 27(11):1455–1461, June 2011.

[8] Jan Schröder, Heiko Schröder, Simon J Puglisi, Ranjan Sinha, and Bertil Schmidt. SHREC: a short-read error correction method. *Bioinformatics*, 25(17):2157–2163, Sep 2009.

[9] Xiao Yang, Sriram P. Chockalingam, and Srinivas Aluru. A survey of error-correction methods for next-generation sequencing. *Briefings in Bioinformatics*, 10.1093/bib/bbs015, 2012.

An investigation of the syntactic flexibility of cis-regulatory modules in ascidians.

Edwin JACOX[1], Mathieu GINESTE[1] and Patrick LEMAIRE[1]

[1] Centre de Recherche de Biochimie Macromoléculaire, UMR5237 CNRS, 1919 route de Mende, 34293, Montpellier, Cedex 05, France

{edwin.jacox, mathieu.gineste, patrick.lemaire}@crbm.cnrs.fr

Keywords Transcription regulation, transcription factor binding, epigenetics, cis-regulatory modules, ascidians.

Ascidian (tunicata) embryos show a remarkable morphological similarity during their development, based on stereotyped and highly conserved cell lineages. We are studying two species in two genera estimated to be 300 million years apart, *Ciona intestinalis* and *Phallusia mammillata*, which in spite of their very similar development have very divergent genomes with almost no sequence alignment other than protein coding DNA. We are investigating if this apparent paradox results from a high level of plasticity in enhancer architecture that would allow extensive transcription factor binding site turn-over without significant consequence on enhancer transcriptional output. To do so, we are identifying early developmental cis-regulatory modules in each species using comparative genomics, and epigenomics. The resulting set of modules are then matched between genera using transcription factor binding site predictions. By comparing the changes in the structure and function of the modules between the two genera, we can investigate the essential aspects of the modules, conserved in the arrangement of transcription factor binding sites.

Cis-regulatory modules are known to be more conserved than the background DNA. Genome assemblies for two *Ciona* species have existed for several years and we have sequenced two closely related species of the *Phallusia* genus. These pairs of genomes allow us to identify potential cis-regulatory modules within the conserved non-coding DNA using comparative genomics. However, these sequences are not conserved between the two genera. To match the modules between the genera, we are using the k-mer composition and the repertoire of transcription factors found in each module to identify homologous cis-regulatory modules.

To further investigate the transcriptional landscape, we are mapping histone modifications associated with cis-regulatory modules (H3K4me3, H3K4me1, and K3K27ac) using ChIP-seq. This data is used to further refine the predicted modules. We are also using this data to examine if the epigenetic landscape is locally conserved between *Ciona* and *Phallusia*. Our initial analysis of the *Phallusia* data indicates that, in addition to intergenic and intronic regions, these marks are often significantly enriched over exons, in contrast to previous work in other species that showed little evidence of these marks on exons.

Using high-throughput Systematic Evolution of Ligands b EXponential enrichment (SELEX), we have determined the in vitro binding affinities for 150 transcription factors of *Ciona intestinalis*. In the SELEX method, a tagged recombinant protein (the transcription factor) is incubated in solution with double-stranded oligonucleotides. Bound oligonucleotides are pulled down, amplified by PCR, and then sequenced by high-throughput methods. The results of each experiment are hundreds to hundreds of thousands of 18- or 20-mers that likely bind the transcription factor. This extensive collection of DNA-binding specificities covers a third of the predicted *Ciona* transcription factors, the largest fraction for any species to date. We find that the motifs are well conserved with vertebrate transcription factors, suggesting that the *Ciona* transcription factor are likely valid in other ascidian species and can also be used to find binding sites in *Phallusia*.

We are using the predicted binding sites, along with an extensive catalog of known expression patterns for most genes in *Ciona*, to identify functional binding sites and to predict the expression pattern of the cis-regulatory modules. These predictions will be verified in vivo. These combined approaches should provide a global view of the degree of cis-regulatory module functional conservation, of the syntactic flexibility of module architecture, and of the conservation of regulatory links between distantly related ascidian species.

GEPETTO: Framework For Gene Prioritization

An open- source framework extended to protein prioritization

Vincent WALTER[1], Hoan NGUYEN[1], Julie D. THOMPSON[1], Olivier POCH[1]

[1] LABORATOIRE DE BIOINFORMATIQUE ET DE GENOMIQUE INTEGRATIVES (LBGI), IGBMC, 1 rue Laurent Friès, 67400 Illkirch-Graffenstaden, France

walterv@igbmc.fr, nguyen@igbmc.fr, julie@igbmc.fr, poch@igbmc.fr

Keywords Gene prioritization, protein prioritization, open-source, framework.

1 Introduction

In the era of omics "big data", and in particular next-generation sequencing (NGS), gene prioritization is a crucial task, involving in the integration of huge amounts of heterogeneous data and the selection and analysis of genes predicted to be involved in a specific biological process, such as pathology. Large sets of genes must be evaluated, in order to score and rank them according to their similarity with known genes and their potential viability as candidates for important applications, such as diagnostic/prognostic markers, drug targets, etc.

2 A new open- source framework for gene prioritization

A customizable and extensible framework is needed for gene selection that can handle large-scale, public and private biological information. There are a lot of gene prioritization applications and web services [1,2] like Endeavour[3,4] or ToppGene[5], but to our knowledge, no other open-source framework for gene prioritization has previously been developed.

GEPETTO (Gene Prioritization Extended Tool) is an original open-source framework, distributed under the LGPL license, for gene selection and prioritization on a desktop computer that ensures confidentiality of personal data. For the time being, it takes advantage of the data integration capabilities in the SM2PH-Central knowledge base [6] (http://decrypthon.igbmc.fr/sm2ph), combined with in-house developed gene prioritization methods. Tomorrow, you will integrate data from your own database developing new modules.

It currently incorporates six prioritization modules 1) gene-based: tissular expression from micro arrays data, disease-causing probabilities using IDGP [7,8], and genomic context, or 2) protein-based: sequence using BLASTp [9,10], protein-protein interactions using STRING-db [11,12], protein evolution using Evolucode[13]. In the future, we intend to extend the system to variant prioritization, by exploiting the variant data in the MSV3D database[14].

GEPETTO is open-source, written in Java/Python and supported by an advanced modular architecture, which means that it can easily be modified and extended by the bioinformatics community, in order to include alternative scoring methods and new public or private data sources.

The software is available at http://sourceforge.net/projects/gepetto/files or as a web service at http://decrypthon.igbmc.fr/sm2ph/cgi-bin/gepetto.

Acknowledgements

The work was performed within the framework of the Decrypthon program, co-funded by Association Française contre les Myopathies [AFM, 14390-15392]; IBM and Centre National de la Recherche Scientifique (CNRS); ANR [Puzzle-Fit: 09-PIRI-0018-02, BIPBIP: ANR-10-BINF-03-02: FRISBI: ANR-10-INSB-05-01];

References

[1] LC. Tranchevent, FB. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker, Y. Moreau. A guide to web tools to prioritize candidate genes. Brief Bioinform. 2011 Jan;12(1):22-32.

[2] Y. Moreau, LC. Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet. 2012 Jul 3;13(8):523-36.

[3] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, Y. Moreau. Gene prioritization through genomic data fusion. Nat Biotechnol. 24(5):537-44, 2006.

[4] L.C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, S. Aerts, Y. Moreau. ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res. 36(Web Server issue):W377-84, 2008.

[5] J. Chen, E.E. Bardes, B.J. Aronow, A.G. Jegga, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 37(Web Server issue):W305-11, 2009.

[6] A. Friedrich, N. Garnier, N. Gagnière, H. Nguyen, L.P. Albou, V. Biancalana, E. Bettler, G. Deléage, O. Lecompte, J. Muller, D. Moras, J.L. Mandel, T. Toursel, L. Moulinier, O. Poch, SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases. Hum Mutat. 32(2):127-35, 2010.

[7] N. Lopez-Bigas, C.A. Ouzounis, Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res. 32(10):3108-14, 2004.

[8] B. Calvo, N. Lopez-Bigas, S.J. Furney, P; Larranaga, J.A. Lozano, A partially supervised classification approach o dominant and recessive human disease gene prediction. Comput Methods Programs Biomed. 85(3):229-37, 2007.

[9] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. J Mol Biol. 215(3):403-10, 1990.

[10] D.J. States, W. Gish, Combined use of sequence similarity and codon bias for coding region identification. J Comput Biol. 1(1):39-50, 1994.

[11] B. Snel, G. Lehmann, P. Bork, M.A. Huynen, STRING: a web-server to retrieve and display the repeatedly occuring neighbourhood of a gene. Nucleic Acids Res. 28(18):3442-4, 2000.

[12] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, L. Simonovic, A. Roth, J. Lin, STRING v9,1; protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res (41(Database issue):D808-15, 2013.

[13] B. Linard, N.H. Nguyen, F. Prosdocimi, O. Poch, J.D. Thompson, Evolutionary Barcodes as a Unifying Framework for Multilevel Evolutionary Data. Evol Bioinform Online, 8:67-77, 2012.

[14] T.D. Luu, A.M. Rusu, V. Walter, R. Ripp, L. Moulinier, J. Muller, T. Toursel, J.D. Thompson, O. Poch, H. Nguyen, MSV3d: database of human MisSense Variants mapped to 3D protein structure. Database (Oxford), 2012.

DBClustalGPU

Alan LAHURE, Olivier POCH and Julie D THOMPSON

Institut de Génétique et de Biologie Moléculaire et Cellulaire, UMR 7104 CNRS/INSERM/Université de Strasbourg, 1 rue Laurent Fries, BP 10142, 67404, Illkirch, Cedex, France

{alan, poch, julie}@igbmc.fr

Abstract *DBClustalGPU is a protein sequence alignment program, based on MSA-CUDA and DbClustal, and developed in C++ and CUDA. MSA-CUDA parallelizes the three steps of the global alignment program, ClustalW (pairwise alignments, guide tree construction and progressive alignment), on GPGPU (General-Purpose Processing on Graphics Processing Units). We have developed DBClustalGPU to integrate local conservation information in MSA-CUDA in the form of a list of anchors. Anchors can be created by any external program, for example the Blast post-processing program, Ballast. DBClustalGPU thus allows faster and more accurate alignment of large sequence sets.*

Keywords Protein sequence alignment, CUDA, GPGPU, anchors

DBClustalGPU

Résumé *DBClustalGPU est un programme d'alignement de séquences protéiques, basé sur MSA-CUDA et DbClustal, et développé en C++ et CUDA. MSA-CUDA parallélise les trois étapes de ClustalW (alignement par paires de séquences, construction de l'arbre guide et l'alignement progressif) sur les GPGPU (General-Purpose Processing on Graphics Processing Units). Nous avons développé DBClustalGPU pour intégrer la conservation locale dans MSA-CUDA en utilisant une liste d'ancres. Ces ancres viennent d'un programme externe, tel que Ballast un postprocesseur de Blast. DBClustalGPU permet un alignement plus rapide mais aussi plus robuste pour un grand ensemble de séquences.*

Mots-clés Alignement de séquences protéiques, CUDA, GPGPU, motifs

1 Introduction

Multiple Sequence Alignment (MSA) plays an important role in the analysis of biological sequences (DNA, RNA or proteins). The introduction of high throughput sequencing technologies has led to a need for faster, more robust MSA tools. The most widely used MSA methods, such as ClustalW [1] and many others, use the progressive alignment algorithm involving three main steps: pairwise distance computation, guide tree construction and profile-profile multiple alignment. Although this is a heuristic algorithm, it is still too computationally expensive for large sets of sequences. One solution to this problem is to parallelize the different steps, either on multiple CPUs or on GPGPUs. For example, MSA-CUDA is a GPU implementation of ClustalW, developed by Liu et al, in 2009 [2].

2 Implementation

To develop DBClustalGPU, we used the source code of MSA-CUDA as a starting point. We then applied an approach similar to that used in DbClustal [3], in order to incorporate local conservation information. The MSAs produced by DbClustal have been shown to be more accurate for complex sequence sets. The input to the program includes a set of sequences to be aligned and a list of weighted ‘anchors’ between pairs of sequences. This list can be obtained from different programs, such as Ballast

[4] (a Blast [5] post-processing program) or [6].

The anchors are incorporated in the third step of the progressive alignment, i.e. during the construction of the profile-profile multiple alignment. During this step, larger and larger sets of sequences, represented by profiles, are aligned in the order defined by the guide tree. In MSA-CUDA, the score for aligning any two profile columns depends only on the values in a residue comparison matrix. We have modified these scores to include the weights of the anchors corresponding to the sequences in each profile. The anchor weights are calculated on the CPU, then loaded into GPGPU memory before the alignment calculation.

3 Results

We compare DBClustalGPU with ClustalW version 2 and DbClustal on CPU and MSA-CUDA on GPGPU. The hardware used is an Intel Core i7-2820QM CPU 2.30GHz and Quadro 1000M (96 cores) and Quadro 5000 (352 cores) GPGPUs. The performances of the different algorithms are evaluated in terms of both execution time and alignment quality using a large set of reference alignments from the BAliBASE benchmark [7]. Figure 1 shows a comparison of the execution time of the DBClustal tool running on either the CPU or the GPU. These tests were performed on six sets of sequences containing a different number and length of sequences:

- set 1 : 11 sequences of average length 230 amino acids
- set 2 : 27 sequences of average length 400 amino acids
- set 3 : 68 sequences of average length 300 amino acids
- set 4 : 75 sequences of average length 500 amino acids
- set 5 : 80 sequences of average length 750 amino acids
- set 6 : 100 sequences of average length 1500 amino acids.

For smaller sets of sequences, the CPU version is faster, due to the time required to transfer the data from the CPU to the GPU and vice versa. For larger sets of sequences, the GPU version is faster.

In conclusion, DBClustalGPU allows to rapidly obtain a reliable alignment, thanks to the parallelisation on GPUs and the integration of locally conserved anchors in the global alignment algorithm.

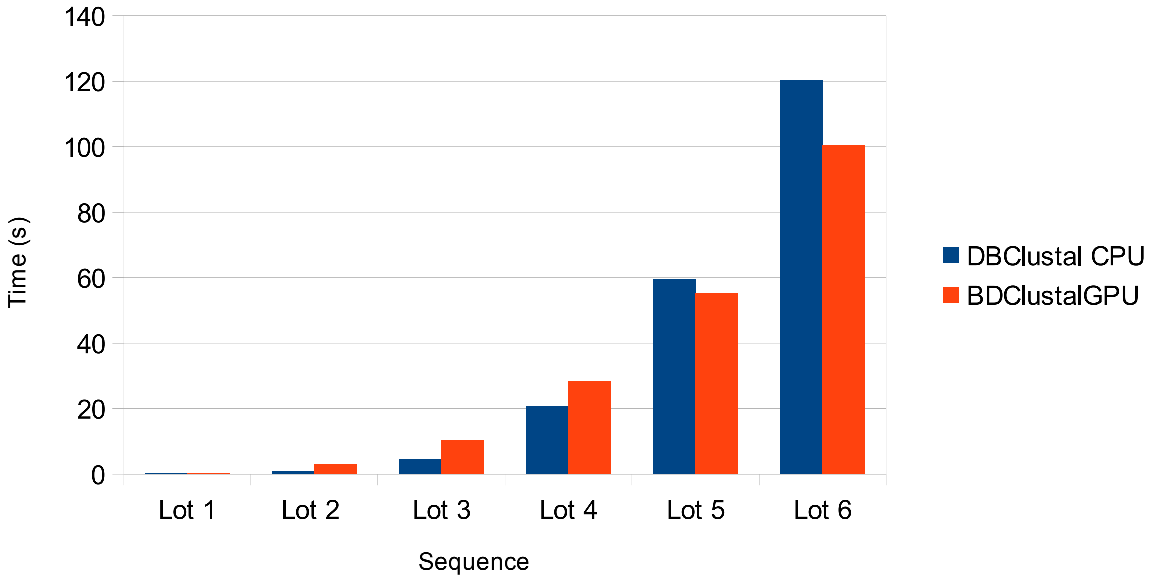


Figure 1. Comparison of execution times for DBClustal running on CPU and GPU

Acknowledgements

We thank Bertil Schmidt and Douglas L. Maskell for the source code of MSA-CUDA and Jean-Christophe Baratault of NVIDIA for the use of the graphics cards.

References

[1] MA. Larkin, G. Blackshields, NP. Brown, R. Chenna, PA. McGettigan, H. McWilliam, F. Valentin, IM. Wallace, A. Wilm, R. Lopez, JD. Thompson, TJ. Gibson and DG, Higgins, Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947-2948, 2007.

[2] Y. Liu, B. Schmidt and DL. Maskell, MSA-CUDA: Multiple Sequence Alignment on Graphics Processing Units with CUDA. *20th IEEE International Conference on Application-specific Systems, Architectures and Processors*, 2009.

[3] JD. Thompson, F. Plewniak, JC. Thierry and O. Poch, DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.*, 28:2919-2926, 2000.

[4] F. Plewniak, JD. Thompson and O. Poch, Ballast: Blast post-processing based on locally conserved segments. *Bioinformatics*, 16:750-759, 2000.

[5] SF. Altschul, TL. Madden, AA. Schäffer, J. Zhang, Z. Zhang, W. Miller and DJ. Lipman, Gapped Blast and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, 25: 3389-3402, 1997.

[6] F. Pitschi, C. Devauchelle and E. Corel, Automatic detection of anchor points for multiple sequence alignment. *BMC Bioinformatics*. 2010 11:445.

[7] JD. Thompson, B. Linard, O. Lecompte and O. Poch, A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One*., 6:e18093, 2011.

A shotgun metagenomic method to identify low abundant species and assign precisely taxonomy in complex microbial ecosystems

Anne-Laure ABRAHAM[1], Mathieu ALMEIDA[1], Nicolas PONS[1], and Pierre RENAULT1[1]

[1] INSTITUT MICALIS, UMR1319 INRA, Domaine de Vilvert, 78352, Jouy-en-Josas, Cedex, France

anne-laure.abraham@jouy.inra.fr, pierre.renault@jouy.inra.fr

Abstract *The extensive characterisation of species present in complex microbial ecosystems requires metagenomic approaches. Most of current approaches are still unreliable for the detection of low abundance species and their precise taxonomy assignation. We present a new tool that gives encouraging results in the characterisation of low abundance species and may allow distinguishing strains under the subspecies taxonomic assignations.*

Keywords Metagenomics, next generation sequencing, microbial genomes.

A first step for a better understanding of complex microbial ecosystems, such as cheese or human gut microbiota, is the characterisation and quantification of their species composition. Once DNA is extracted from samples, two main techniques can be used: the sequencing of specific phylogenetic markers, such as genes coding for the 16S or 18S RNA or ITS, or whole genome shotgun sequencing. The first approach is widely used and provides a rapid view of the ecosystem, but often fails to provide precise taxonomy information due to sequencing errors and biased quantification due to amplification issues. Shotgun metagenomic sequencing approaches may circumvent such issues. They are often based on the use of gene marker sets to discriminate species without ambiguity, although the use of a small part of the genomes decreases sensitivity of this approach.

We developed a new approach in order to characterise precisely the taxonomy of the present species, up to the strain levels in some cases, and to identify low abundance species. For this purpose, the metagenomic reads are mapped on whole genomes from a dataset of reference strains. Then, for each reference genome, we use the Lander and Waterman equation [1] to compute the proportion of the genomes that is expected to be covered by reads if they were distributed randomly across the genome. The comparison of the expected and the observed values allows to identify the present species and filter false positive results, especially for low abundant species which display a small number of reads. In a second step, we measure strains divergence to the reference genomes and perform taxonomical assignment. Additional information such as gene composition and SNP counting may allow assessing up to strain level.

We tested this tool on several datasets, including simulated reads and reads from genomic and metagenomic sequencing projects in order to test the power and limits of this method. We will present examples on cheese ecosystems.

Acknowledgements

This work was supported by ANR FOOD-MICROBIOMES.

References

[1] ES. Lander, MS. Waterman, Genomic mapping by fingerprinting random clones: a mathematical analysis, *Genomics* 2(3): 231-239, 1988.

Visual Analytics of Molecular Simulation in Immersive Environment

Mikael TRELLET[1], Nicolas FEREY[1] , Marc BAADEN[2] and Patrick BOURDOT[1]

[1] LIMSI, UMR3251 CNRS, rue John Von Neumann, 91400, Orsay, France
{mikael.trellet, nicolas.ferey, patrick.bourdot}@limsi.fr
[2] LBT - IBPC, UMR9080 CNRS, 13 rue Pierre et Marie Curie, 75005, Paris, France
marc.baaden@ibpc.fr

Abstract *In high performance molecular simulation, data handling and storage became lately a crucial issue considering the available production rate recent computers can offer. The data that a user has to get back from his simulation often reach the critical storage limit for the largest systems. To deal with this technical problem, In Situ approach provides a first solution aiming to reduce the quantity of data laboratories have to store after a simulation. In this approach, analyzes and visual rendering are mainly performed in the distant calculation sites, directly from the raw data produced by a simulation. However, planning of such analyzes is difficult and requires a long decision process upstream. Our approach tackles this planning issue by enabling a total interactivity between the distant running simulation and the users. Users will be able to interactively launch particular analyzes on a running simulation and see the direct interpretation of their analyzes in a graphical space. This approach is facilitated by a first implementation in an immersive environment with a large display for the two main spaces as well as natural navigation and interaction that immersion naturally offers.*

Keywords Visual analytics, molecular simulation, In Situ simulation

1 Introduction

Dealing with large molecular systems and/or long time simulation raise some crucial data storage issues for both laboratories and distant calculation sites. Nowadays, computers performances allow scientists to perform simulation of duration and size such important that they generate up to several To of data. Beyond this storage issue, we know that relevant information a scientist computes from a set of simulation data is quite negligible compared to the few To he stored. Several approaches try to address these issues by either reducing the amount of data generated by a simulation, but reducing also the number of information available for a user, or giving the possibility to perform sone analysis or rendering steps while a simulation is running, avoiding then the storage of raw data already analyzed during the simulation loop. The In Situ approach [1] fits in the second category and aims to identify relevant analyzes a user might want to be performed on his simulation. This approach provides methods and tools that can be applied on a running simulation in order to extract useful information. Unfortunately, this approach requires a solid and accurate planning of the analyzes that will be performed and asks for the user to keep a certain degree of anticipation to know where his simulation will head.

In order to tackle this issue, we propose to add a total interactivity between the user and a running simulation. Instead of planning in advance what analyzes a user might need to perform, we let him the possibility to drive himself the information he needs in function of the current state of his simulation. By enabling a comfortable navigation in the system simulated and the possibility to display live analyzes on this system, we put the user in the center of a decision process . This approach aims to efficiently shorten the data processing flow by gathering the simulation and analytical time together. To do so, we will use the tools and methods provided by the new field of “Visual analytics” [2].

2 Visual Analytics

The quick emerging field of Visual Analytics aims to facilitate analyzes on large and complex data using two main approaches: A complexity reduction of the data visual representation as well as strong interactive visual interfaces. Size, complexity and the need of a strong interaction between human and machine makes this field important to certain problems that regular approaches cannot afford. Visual analytics groups tools coming from several fields and try to marry them to get a solid process of analytic reasoning. Data representation and interactive capability between analyzes and graphical rendering are the main axes developed in visual analytics and the ones we are interesting in.

In the field of molecular simulation, numerous information can be used to represent and analyze a molecule evolution upon time. Several physical or chemical factors, like energies, temperature or distance, come and evolve during a simulation. To fit the visual analytics approach, we need to setup a clear representation of the simulation data, both from the graphical point of view than the analytical one. This representation will take place through a conceptual scheme that will semantically identify each element we will have to manipulate. Moreover, the scheme will also takes into account any interaction and path a data can overcome in order to facilitate their binding between the graphical and the analytical spaces in the immersive environment.

A example can be simply illustrate and the figure 1 reports the basic feature we could find in our final implementation of the approach. In this figure, a 3D chart displays in real time the evolution of the charge and the distance of each chain of a given protein. In parallel, a 3D representation of the protein is shown, highlighting the conformational changes overtaken by the protein during the running simulation. The user will have the possibility to select an analytical element like the edge of one of the displayed curve and then see the corresponding graphical element highlighted by the rendering module. The opposite path is of course possible and a selection of a protein element like an atom or a residue can restraint or highlight a part of a running analysis and will propose to the user a bench of relevant analyzes he would want to perform.

The capacity to display interactive analysis and rendering it in real time is allowed by the component-based approach used to develop FlowVR [3], the program on which we develop our project. FlowVR already shows wonderful performances in the In Situ field by providing the possibility for a user to drive simulations in real time. We want to extend FlowVR performances with analyzes and immersive modules, bringing new possibilities for its futures applications. The immersive part uses a CAVE-like system, EVE, developed in our laboratory (*www.**limsi**.fr/venise/**EVE**system*) and has been optimized for a smooth and natural navigation and an efficient interaction with the molecular structure and the analytical part. The large display capacity of the system let us the possibility to target very large molecular systems such as whole ribosomes, virus or transmembrane protein with explicit membranes and to associate many analyses in the same time.

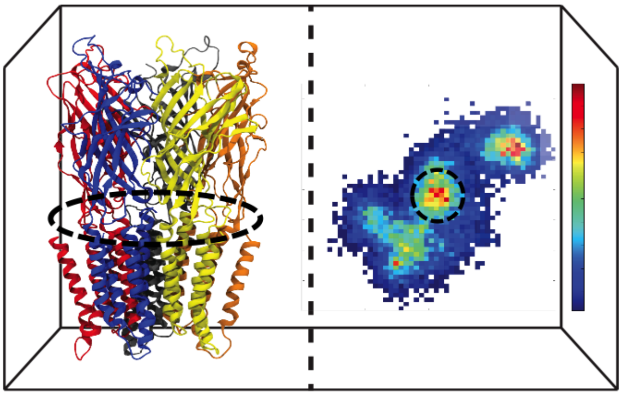


Figure 1. Example illustration of the structural (left) and analytical (right) parts of a molecular simulation in EVE. The surrounded elements are semantically linked and selection of one triggers actions at the second one level.

References

[1] B. Whitlock, J.M. Favre, J.S. Meredith, Parallel In Situ Coupling of Simulation with a Fully Featured Visualization System, *Eurographics Symposium on Parallel Graphics and Visualization*, pp. 101-109, 2011.

[2] J. Kielman, J. Thomas, R. May, Fondation and Frontiers in Visual Analytics, *Information Visualization,* pp. 239-246, 2009.

[3] J-D. Lesage, B. Raffin, High performance interactive computing using FlowVR, *Proceedings of IEE Virtual Reality SEARIS Workshop*, pp. 13-16, 2008.

UMD-Predictor v2

A novel mutation pathogenicity prediction tool.

Jean-Pierre Desvignes[1,2] & David Salgado[1,2] Ghadi Raï[1,2] ,Gwenaelle Collod-Beroud[2] and Christophe Béroud[1,2,3]

[1] INSERM, UMR_S 910 , Faculté de Médecine de la Timone, 27 Bd Jean Moulin, 13385, Marseille, Cedex 05, France

[2] AMU, Faculté de Médecine de la Timone, 27 Bd Jean Moulin, 13385, Marseille, Cedex 05, France

[3] APHM, Laboratoire de Génétique Moléculaire, Hôpital Timone Enfants, 8ème étage, 264 rue Saint Pierre, Marseille, Cedex 05, France

Keywords Prediction, pathogenicity, mutation, SNP

1 Introduction

Non-synonymous single nucleotide polymorphism (nsSNP) in the coding region of genes causing amino acid substitution may have a large impact on protein function. With next generation sequencing, the possibilities to identify nsSNP have increased. Distinguishing neutral sequence variations from those responsible for the phenotype is a major interest in human genetics.

To further differentiate neutral variants from pathogenic nucleotide substitutions, we developed a tool, UMD-Predictor. This tool provides a combinatorial approach that associates the following data: localization within the protein, conservation, biochemical properties of the mutant and wild-type residues, and the potential impact of the variation on mRNA. UMD-Predictor v2 is an adaptation of UMD-Predictor [1] with a fully redesigned data management system, the creation of a web interface and of a webservice.

2 UMD-Predictor v1

UMD-Predictor v1 is implemented on 4D platform (SGBDR) and is based on hg18 version of the human genome. 24 databases representing each chromosome were developed, each of one integrates various sources of annotation and algorithms allowing calculation of pathogenicity prediction of the mutations (Fig 1). This version has been evaluated with a 796 mutations dataset based on 4 genes. With this dataset, our tool has shown a positive sensitivity of 95,4 % and a specificity of 92,2 % demonstrating that the combinatorial approach used by UMD-Predictor gives better results than other tools [1].

3 UMD-Predictor v2

3.1 Changes

The newly developed version of UMD-Predictor is based on the hg19 version of human genome and on Ensembl v69 gene annotations. We have ported all 4D databases into one optimized PostgreSQL database for all chromosomes. A web site is now available with an user-friendly interface. Several options are available to users. End users can recover all predictions for an entire transcript using a GeneSymbol, an Ensembl Transcript Id or a refseqID. Users can provide as input a list of mutations in various format directly on the website or by uploading a file (txt or vcf). We have also developed a webservice to allow pragmatic access to all predictions available into our tools. More than 268 millions of subtitutions was precomputed versus 179 millions for the v1.

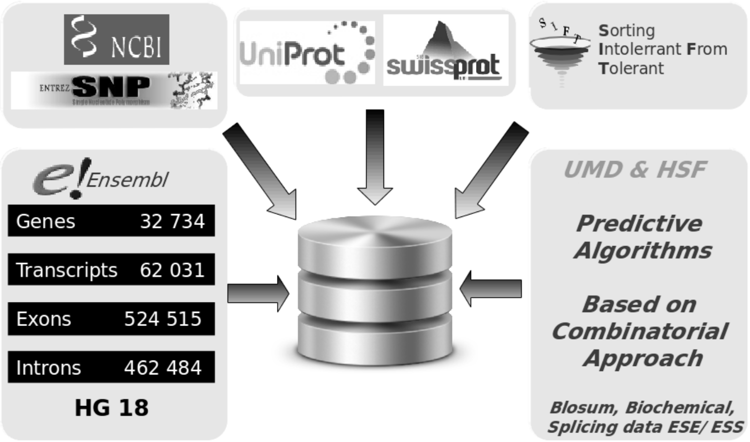


Figure 1. Data integrated into UMD-Predictor v1.

3.2 Evaluation

The evaluation of UMD-Predictor v2 is currently in progress (in date of submission). We are evaluating our software against 5 prediction tools (SIFT, Polyphen2, Mutation Taster, Mutation Assessor, Condel). The dataset includes 16248 pathogenic mutations (from varibench dataset[2]) and 7532 neutral mutations (from DbSNP 135 with frequency > 1 % and number of individu > 50).

4 Conclusions

The prediction of the pathogenic impact of a given missense mutation is one of the biggest challenges for human genetics and molecular diagnostic laboratories.

In the past whe have demonstrated that UMD-Predictor v1 was an accurate genetic mutation prediction tool. Some limitations such as the lack of a dedicated web interface and old data (HG18) made the proposed update mandatory. We hope with this update, and improvements in progress, to obtain better results and offer to geneticists and other scientists an efficient tool adapted to current diagnostic and research strategies such as exome or whole genome sequencing.

Acknowledgements

This work was supported by RD-CONNECT.

References

[1] MY. Frederic, M. Lalande, C. Boileau, D. Hamroun, M. Claustres, C. Beroud and G. Collod-Berroud, UMD-Predictor, a New Prediction Tool for Nucleotide Substitution Pathogenicity. Application to Four Genes: FBN1, FBN2, TGFBR1, and TGFBR2 . *Hum Mutat.,* Jun;30(6):952-9, 2009

[2] P. Sasidharan Nair P, M. Vihinen , VariBench: a benchmark database for variations. Hum Mutat, Jan;34(1):42-9, 2013

Human Splicing Finder 3.0

New release of an online bioinformatics tool predicting splicing signals

Ghadi Raï[1], David Salgado[1,2], Gaëlle Blandin[1], Jean-Pierre Desvignes[1], Gwenaëlle Collod-Béroud[1] and Christophe Béroud[1,3,4]

[1] INSERM, UMR_S 910 , Faculté de Médecine de la Timone, 27 Bd Jean Moulin, 13385, Marseille, Cedex 05, France

[2] Australian Regenerative Medicine Institute, EMBL Australia, Monash University, Building 75, Clayton, Victoria 3800, Australia

[3] AMU, Faculté de Médecine de la Timone, 27 Bd Jean Moulin, 13385, Marseille, Cedex 05, France

[4] APHM, Laboratoire de Génétique Moléculaire, Hôpital Timone Enfants, 8ème étage, 264 rue Saint Pierre, Marseille, Cedex 05, France

Keywords: Splicing, bioinformatics, HSF, prediction, variation, mutation, human.

1 Introduction

With the completion of the Human Genome Project our vision of human genetic diseases has changed. Yearly, thousands of mutations are identified in diagnostic and research. Although many of them directly affect protein expression, it has been demonstrated that a large number of mutations influence mRNA splicing. Existing splice sites are mostly affected, but mutations can also create or disrupt splice sites or auxiliary *cis*-splicing sequences.

Human Splicing Finder [1] identifies splicing motifs in any human sequence and predicts the effects of a mutation on these signals. New algorithms were developed to determine consensus values of potential splice sites and search for branch points. Furthermore, all available matrices have been integrated to identify exonic and intronic motifs, as well as matrices to identify *hnRNP A1*, *Tra2-β* and *9GB*. On this new version, we have updated the database content, developed a new web application, and implemented new tools to assist users to interpret their results.

2 About Human Splicing Finder

Human Splicing Finder is currently one of the more accurate online free tools in the field of splicing prediction. It has been cited more than 250 times over the past three years. The 2.4.1 version was developed using the 4D v11 software (SGBDR), and its data was based on the hg18 version of the human genome and the Ensembl database (release 54). Its algorithms are based on Position Weight Matrices, to assess the strength of 5' and 3' splice sites, branch points and ESE/ESS auxiliary motifs. HSF also includes an algorithm adapted from the MaxEnt script [2] that allows the analysis of a whole sequence.

3 What's new on HSF v3.0

3.1 Updates and upgrades

As a first step, we upgraded HSF to the 4D - v13 software in order to benefit from the new functionalities offered by the platform. Concomitantly, the content of the database has been updated with the latest human genome version hg19, the release #70 of the Ensembl database [3] and the latest release of the HGNC database [4]. In addition, we developed a new web application with a user-friendly interface, allowing analysis using various types of data as entry points and results visualization in tabs or/and in dynamic graphics.

3.2 New adding and tools

In order to add another point of comparison of results, we integrated the NNSplice [5] prediction algorithm in the "analyze a sequence" part of HSF. Furthermore, we are computing new matrices to predict "non *GT-AG"* introns, which represent less than 2% of human introns, in order to cover the whole human genome. Additionally, we are actively working on the development of the "help to interpret" tool. This tool will guide users through their "analyze a mutation" results, predicting its impact on splicing (exon skipping or use of cryptic splice sites).

4 Conclusions

Detection of splicing motifs and prediction of mutation impact on splicing signals is a major topic in the field of bioinformatics and genetics. Human Splicing Finder has proven to be one of the most widely used tools to evaluate the impact of mutations on splicing signals. As the Human Genome reference sequence is periodically updated and new scientific knowledge is available for splicing, we created a fully revised version of HSF. Improvements in progress should, hopefully, enhance its accuracy in order to better help scientists and geneticists in their research and diagnostic day-to-day practice.

Acknowledgements

This work was supported by the GEN2PHEN project. We also want to thank all users who demonstrate the interest and efficiency of HSF.

References

[1] F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, C. Béroud, Human Splicing Finder: an online bioinformatics tool to predict splicing signals, Nucleic Acids Research, 37 (2009) e67.

[2] G. Yeo, C.B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, J. Comput. Biol, 11 (2004) 377–394.

[3] P. Flicek, I. Ahmed, M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. Garcia-Giron, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A.K. Kahari, S. Keenan, M. Komorowska, et al., Ensembl 2013, Nucleic Acids Research, 41 (2012) D48–D55.

[4] K.A. Gray, L.C. Daugherty, S.M. Gordon, R.L. Seal, M.W. Wright, E.A. Bruford, Genenamesorg: the HGNC resources in 2013, Nucleic Acids Research, 41 (2013) D545–52.

[5] M.G. Reese, F.H. Eeckman, D. Kulp, D. Haussler, Improved splice site detection in Genie, J. Comput. Biol, 4 (1997) 311–323.

TriAnnot: a powerful bioinformatics tool for plant genome gene identification

Nicolas GUILHOT[1], Sébastien THEIL[1], Frédéric CHOULET[1], CHRISTOPHE CARON[2], ALEXANDRE CORMIER[3], Catherine FEUILLET[1] et Philippe LEROY[1],

[1] INRA-UBP, UMR 1095 Génétique, Diversité and Ecophysiologie des Céréales, 5 chemin de Beaulieu, F-63039 Clermont-Ferrand cedex2, France

philippe.leroy@clermont.inra.fr

[2] CNRS, UPMC, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France

[3] UPMC, CNRS, UMR7139, Génétique des Algues, Station Biologique, 29680, Roscoff, France

Keywords bread wheat, structural and functional annotation, genes, ncRNAs, transposable elements, molecular markers, plant genomes, pipeline, cluster.

1. CONTEXTE ET ENJEUX

Les caractères agronomiques tels que rendement, résistances aux maladies ou aux facteurs environnementaux sont déterminés par l'information codée par les gènes et éléments régulateurs portés par les chromosomes des plantes. Les capacités adaptatives des plantes aux changements environnementaux reposent en grande partie sur l'utilisation des variations de cette information, soit par les mécanismes de la sélection naturelle (évolution et spéciation), soit via des programmes d'amélioration génétique dirigés par les sélectionneurs (progrès génétique). Décrypter par séquençage la molécule d'ADN constituant les chromosomes fournit ainsi des informations essentielles pour mieux comprendre et améliorer les caractères d'intérêt en sélection variétale. En effet, l'un des enjeux majeurs de la génomique végétale est de fournir des connaissances et des outils performants pour la sélection, sur un pas de temps court, de nouvelles variétés cultivées à même de répondre aux enjeux de la sécurité alimentaire mondiale dans de nouvelles conditions environnementales générées par le changement climatique, et à travers des pratiques culturales plus respectueuses de l'environnement.

Un séquençage de haute qualité permettant d'avoir une information la plus complète possible ordonnée le long des chromosomes est un pré-requis à tout projet d'annotation structurale et fonctionnelle. Ensuite, une annotation rigoureuse et fiable est essentielle pour une utilisation efficace de l'information pour la recherche des gènes candidats, l'analyse de polymorphismes et la recherche de nouveaux allèles, des études des réseaux de régulation et des études sur l'évolution et la génomique comparée avec d'autres espèces. De plus, une mauvaise annotation peut polluer les bases de données et compromettre par la suite l'annotation des génomes d'autres espèces végétales. L'annotation consiste à prédire à partir des séquences la structure et le rôle potentiel de gènes, d'éléments transposables et d'autres séquences non codantes pour des protéines mais pouvant avoir un rôle dans la régulation de l'expression des gènes. Ces prédictions sont basées sur l'utilisation de programmes informatiques et de comparaisons avec des bases de données. L'annotation des génomes ne peut se faire entièrement manuellement et la possibilité de réaliser une première analyse, de façon automatisée pour prédire de façon la plus robuste et fiable possible les annotations, représente un avantage majeur dans les projets de séquençage des génomes. Avec le séquençage prévu des 21 chromosomes du génome de blé hexaploïde (17Gb), l'*International Wheat Genome Sequencing Consortium* (IWGSC - http://www.wheatgenome.org/) avait besoin d'un outil performant. Le projet de séquençage du chromosome 3B porté par l'INRA a permis de mettre en place et évaluer les performances d'un pipeline automatisé, TriAnnot (http://www.clermont.inra/triannot).

2. RESULTATS

L'architecture du pipeline TriAnnot est modulaire. TriAnnot permet d'annoter et/ou de masquer les éléments transposables; de proposer une structure et une fonction biochimique pour les gènes codant pour des protéines avec un index de qualité basé sur des évidences biologiques; d'identifier des séquences non-codantes conservées et de définir des marqueurs moléculaires. Le pipeline est parallélisé sur un cluster mettant à disposition environ 900 cœurs, et peut annoter automatiquement un génome/chromosome de 1 Gb en moins de 2 jours. Le pipeline est accessible en ligne grâce à une interface web conviviale pour des analyses de quelques centaines de séquences (BAC, Scaffold de 10 kb à 3 Mb), mais il peut être également utilisé directement sur le cluster de l'URGI à Versailles grâce à un programme spécial de soumission écrit en PERL pour des analyses à l'échelle de chromosomes ou génomes entiers.

Le pipeline TriAnnot a été dans un premier temps évalué avec des séquences de blé de référence en comparaison d'outils d'annotation existants. Il a montré une plus grande fiabilité que des pipelines non optimisés pour le blé. Il a également été testé sur le chromosome 1 de riz et a permis de détecter des annotations erronées. TriAnnot a permis d'annoter les 1 Gb de séquence du chromosome 3B (deux fois et demie le génome de riz) en 26 heures. Depuis sa publication [1], des améliorations importantes ont été réalisées en particulier sur la validation automatisée des prédictions de gènes ce qui réduit considérablement le travail de curation manuelle toujours nécessaire dans les projets de séquençage de génome. TriAnnot peut désormais distinguer les gènes complets des gènes fragmentés et des pseudogenes avec une note de confiance (HC – High Confidence ; LC – Low Confidence). TriAnnot est également utilisé actuellement pour l'annotation de contigs d'orge (IPK, Allemagne), de BAC de chêne (INRA URGV, Evry) et de scaffolds du chromosome 4D de blé (Collaboration avec INTA, Argentine). Un site web a été mis en place et la version 3.8 est en ligne (http://www.clermont.inra.fr/triannot).

3. PERSPECTIVES

Les perspectives d'amélioration sont de (1) construire une Machine Virtuelle du pipeline en collaboration avec la plateforme ABiMS de Roscoff et l'Université de Saskatchewan au Canada pour permettre a d'autres équipes d'utiliser le pipeline localement, (2) améliorer l'architecture du pipeline et ajouter des modules pour les petits ARN (RNAspace développé par l'Equipe de C. Gaspin à l'INRA de Toulouse) et d'autres marqueurs moléculaires à haut débit et (3) valider un projet pilote avec le VIB (Gand, Belgique) afin d'interfacer le pipeline avec un environnement graphique (ORCAE) pour l'expertise manuelle de l'annotation automatique en ligne dans le cadre de l'appel d'offre à projet LifeGrid 2012 (Conseil Régional d'Auvergne/FEDER). Enfin, le code de TriAnnot sera modifié de manière à permettre l'annotation, par fenêtre glissante, d'une pseudomolécule d'environ 1Gb.

4. VALORISATION

TriAnnot a pour ambition de fédérer la communauté internationale autour de l'annotation des 21 chromosomes de blé tendre dans le cadre de la *wheat Initiative* (http://www.wheatinitiative.org/) et de l'IWGSC et de fournir un outil à d'autres communautés faisant face à des défis similaires. Il est utilisé par l'équipe Génome de l'Unité INRA-UBP GDEC de Clermont-Ferrand qui est en charge des chromosomes 3B et 1B et sera utilisé prochainement pour l'annotation des chromosomes 6B (Japon) ; 1A (Canada) et 4D (Argentine), et du génome d'*Aegilops tauschii* (ancêtre du génome D de blé tendre) dans le cadre d'un projet NSF américain coordonné par J. Dvorak.

Références

[1] P. Leroy, N. Guilhot, H. Sakai, A. Bernard, F. Choulet, S. Theil, S. Reboux, N. Amano, T. Flutre, C. Pelegrin, H. Ohyanagi, M. Seidel, F. Giacomoni, M. Reichstadt, M. Alaux, E. Gicquello, F. Legeai, L. Cerutti, H. Numa, T. Tanaka, K. Mayer, T. Itoh, H. Quesneville, C. Feuillet, TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. *Frontiers in Plant Sciences,* 3:1-14, 2012

Design and Development of Galaxy Workflows for Microbiome Analyses

Arnaud FELTEN[1], Olivia DOPPELT-AZEROUAL[2], Fabien MAREUIL[2],
Pierre DEHOUX[1], Corinne MAUFRAIS[2], Catherine DAUGA[1]
[1] DEPARTMENT OF GENOMES AND GENETICS, INSTITUT PASTEUR, 28 rue du Docteur Roux - 75015 Paris, France
{arnaud.felten, pierre.dehoux, catherine.dauga}@pasteur.fr
[2] CENTRE OF INFORMATICS FOR BIOLOGY, INSTITUT PASTEUR, 28 rue du Docteur Roux - 75015 Paris, France
{corinne.maufrais,olivia.doppelt,fabien.mareuil}@pasteur.fr

Keywords : Microbiome, Galaxy Workflows, GemSim, 16S rRNA

1. Introduction

Distinct microbial communities inhabit the body's surfaces: the skin, the oral cavity, the upper respiratory tract, the genital tract and the stomach of humans. It is very interesting to search associations between microbiome structure and clinical phenotype in human diseases by showing significant levels of variation of some microbial components.

Many past and present metagenomic studies generally do not address the effect of classification approaches on the accuracy of the results, even though it is of great significance for microbiome research. For instance, the bacterial component of the human gut microbiome consists of ≥1,800 genera and an astonishing 15,000–36,000 species, depending on whether species are classified conservatively or liberally [1]. We thought it important to take in consideration the performance of approaches inferring the taxonomic origin of reads.

Probably the most fast and simple method is to use BLAST to search for the best hit or for several hits classified by the lowest common ancestor (LCA, [2]) approach, to provide taxonomic information of detected homologies. We have chosen the Galaxy platform [3], with its user-friendly interface, to develop workflows testing the efficiencies of classification obtained by homology-based approaches, and to adapt tools and parameters accordingly. At end, biologists can design a workflow adapted to the nature of the flora and/or to the pathogens to detect, to carry out the taxonomic assignment of their real microbiome data.

2. Limitations of existing tools

We compared classifications of reads from pyrosequencing (454 Titanium) and Illumina technologies. Collections of reads corresponding to complete 16S rRNA and its V5-V6 extended regions with the sequencing simulator called GemSIM were used [4]. We generated synthetic datasets of unique and multiple bacterial species by mimicking the composition of real data obtained from intestinal tract of mice. BLAST search of the reads for each synthetic community was conducted against four 16S rRNA databanks, Greengenes, Silva, RDP2 and the 16S microbial rRNA database (NCBI), currently used in metagenomics.

When comparing classification efficiencies, we observed better assignments of reads for 454, and more unidentified and misidentified reads for Illumina's technology. The choice of tools and parameters and numerous biases affect the accuracy of taxonomic assignments for both technologies:
- The use of next-generation sequencing to perform 16S rRNA sequence surveys has resulted in important controversy surrounding the effect of sequencing errors on downstream analysis [5][6]. We showed that the trimming step decreases false identifications but also can reduce the diversity of microbiota.
- We found that databanks have a significant impact on the result of metagenomic analyses. Silva and Greengenes seem to generate the more reliable identifications. Some bacteria of medical interest, such as *Akkermansia muciniphila* and candidatus *Arthromitus*, are detected more rarely than expected. The low number and/or sequencing and annotation errors of sequences in databanks may explain this bias.
- In comparing the BLAST best-hit and the Lowest Common Ancestor (LCA) approaches, we showed that LCA improves the specificity of results but lacks of sensitivity.

- The taxonomic assignments of the data vary widely according to the clustering method used. The clustering step, in reducing the BLAST search to one representative sequence by cluster, decreases the computation time. Unfortunately many clusters are not well defined and the taxonomic classification of sequence reads differs quite substantially from the original composition of the microbiome.

The results of this evaluation suggest to adapt the strategy of bioanalysis according to *a priori* knowledge of the taxonomic composition of the microbiota and to the biological question asked by the study.

3. Design of three Galaxy workflows to improve taxonomic assignment of microbiomes

3.1 Workflow for Simulations

This workflow allows to assess classification efficiencies of strategies included in the workflow for microbiome analysis (see 3.3), in using synthetic data sets. Biologists can build their own empirical error models from real data by using GemErr, or use default error models of GemSIM to generate 454 or illumina reads (FASTQ) from 16S rRNA reference sequences with known taxonomy. Classifications of both the simulated reads and their reference sequences are compared to highlight qualitative and quantitative biases.

3.2 Workflow for Enrichment of Databanks

If one or more bacteria are lacking, we propose to submit a multi-fasta file of the 16S rRNA from these bacterial species through the workflow for enrichment to improve databanks.

After this process, tools, parameters and enriched databanks are available to build the workflow of analysis.

3.3 Workflow for Microbiome Analysis

This workflow proposes in option, two different trimming processes according to the sequencing technology used and two clustering programs, CD-HIT [7] or CROP [8]. The BLAST search can be performed for all the reads or for one read per cluster. Then, biologists can favor the best-hit approach or the LCA strategy. Taxonomic identification is achieved by taxoptimizer[1] and representation is generated by rankoptimizer[2] with Krona-2.0 library [9].

The 3 workflows will be made available on Galaxy at the Institut Pasteur with a parallelized version of BLAST 2.2.21 (pblastall) and taxoptimizer (ptaxotptimizer) (C. Maufrais, unpublished).

[1] http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::taxoptimizer
[2] http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::rankoptimizer

References

[1] MJ Pallen. The Human Microbiome and Host interactions, in Karen E. Nelson (ed), *Metagenomics of the Human Body*. pp. 43-61, 2011

[2] A Aho, J Hopcroft, J Ullman. On finding lowest common ancestors in trees. *Proc. 5th ACM Symp. Theory of Computing (STOC)* pp. 253–265, 1973

[3] J Goeck, A. Nekruenko, J. Taylor and the Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 25;11(8):R86, 2010

[4] KE McElroy, F Luciani, T Thomas. GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics* vol. 13 pp. 74, 2012

[5] MJ Claesson et al. Comparative Analysis of Pyrosequencing and a Phylogenetic Microarray for Exploring Microbial Community Structures in the Human Distal Intestine. *PLoS ONE* vol. 4 (8) pp. e6669, 2009

[6] PD Schloss, D Gevers, SL Westcott. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* vol. 6 (12) pp. e27310, 2011

[7] L Weizhong, A Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* vol. 22 (13) pp 1658-1659, 2006

[8] X Hao, R Jiang, T Chen. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* doi: 10.1093/bioinformatics/btq725, 2011

[9] Ondov BD, Bergman NH, and Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011 Sep 30; 12(1):385

A Web Site for detecting protein structural domain neighbours

Franck Samson[1], Richard Shrager[2], Jean-François Gibrat[1], Chin-Hsien Tai[3], Vichetra Sam[2], Peter J.Munson[2] and Jean GARNIER[1,2]

[1] Mathématique, Informatique et Génome (MIG) , UR1077 INRA, Domaine de Vilvert, 78350, Jouy-en-Josas, Cedex, France
{franck.samson, jean-francois.gibrat, jean.garnier}@jouy.inra.fr

[2] Mathematical and Statistical Computing Laboratory, CIT, NIH, Bethesda, MD, 20892, USA
{munson, shragerr, vichetra.sam}@mail.nih.gov

[3] Laboratory of Molecular Biology, NCI, NIH, Bethesda, MD, 20292, USA
{taic}@pop.nci.nih.gov

Keywords Protein, Structure, Alignment, Recurrence.

A query protein structure is compared with the VAST program to a database of target structures from the PDB (PDB40, list of protein structures having less than 40% of identical residues: 19 500 structures version 2011). The threshold of the VAST program is lowered in order to find the largest possible number of structures having a local similarity with the query protein. The purpose of the web site is to define structural domains in the query protein using the recurrence of these locally similar substructures (http://genome.jouy.inra.fr/domire/). The list of matches is subsequently sorted according to two criteria: the number of aligned residues by VAST is at least 40% of the number of residues of the target, and 80% of the target length is aligned including gaps of non aligned residues if less than 40. Besides this list, a residue-residue alignment of the structural neighbour on the amino acid sequence of the query protein is provided together with a 3D view of their superposition. The object of this sorting is to help in detecting remote homologues and isolated protein structures matching the domain structures of a protein [1] .

References

[1] F. Samson, R. Shrager, C-H. Tai, V. Sam, B L. Peter, J. Munson, J-F. Gibrat and J. Garnier, Domire: a web server for identifying structural domains and their neighbors in proteins. *Bioinformatics*. 28 (7) 1040-1041, 2012

BIOS: a BioInformatics Oriented Service architecture for RNA-seq analysis

http://bios.toulouse.inra.fr

Sébastien CARRERE [1], Emmanuel COURCELLE [1], Marion VERDENAUD [1], Eric BIOT [2], Erika SALLET [1], Emeline DELEURY [3], Loic LEDANTEC [4], Cécile FIZAMES [5], Jean-Pierre GAUTHIER [6], Vincent SAVOIS [7], Susete ALVES-CARVALHO [7], Philippe GREVET [8], Véronique BRUNAUD [8], Fabrice LEGEAI [6], Bernhard GSCHLOESSL [9], Virginie GARCIA [10] and Jérôme GOUZY [1].

[1] Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR INRA-CNRS 441/2594, F-31320 Castanet Tolosan, France
{Sebastien.Carrere, Emmanuel.Courcelle, Erika.Sallet, Jerome.Gouzy}@toulouse.inra.fr

[2] Institut Jean-Pierre Bourgin, UMR1318, INRA-AgroParisTech, Versailles, France
Eric.Biot@versailles.inra.fr

[3] Intéractions Biotiques et Santé Végétale (IBSV), INRA UMR/CNRS 1301/6243, F-06903 Sophia-Antipolis, France.
Emeline.Deleury@sophia.inra.fr

[4] Unité de Recherches sur les Espèces Fruitières (UREF), F-33883 Villenave d'Ornon, France
Loic.LeDantec@bordeaux.inra.fr

[5] Institut de Biologie Intégrative des plantes, UMR 5004-CNRS/0386-INRA/SupAgro/Univ. Montpellier 2, F-34060 Montpellier, France.
Cecile.Fizames@supagro.inra.fr

[6] BIO3P, UMR1099 INRA/Agrocampus Rennes/Univ. Rennes I
{Jean-Pierre.Gauthier, Fabrice.Legeai}@rennes.inra.fr

[7] INRA, UMR 102 Génétique et Ecophysiologie des Légumineuses, F-21065 Dijon, France
Vincent.Savois@dijon.inra.fr

[8] Unité de Recherche en Génomique Végétale (URGV), UMR INRA 1165 – Univ. d'Evry Val d'Essonne - ERL CNRS 8196, F-91057 Evry, France
{Philippe.Grevet, Veronique.Brunaud}@evry.inra.fr

[9] Centre de Biologie et de Gestion des Populations (CBGP), UMR INRA-IRD-CIRAD-SupAgro, F-34988 Montferrier/Lez, France
Bernhard.Gschloessl@supagro.inra.fr

[10] Biologie du fruit et Pathologie (BFP), UMR1332 INRA-Univ. Bordeaux I&II, F-33882, Bordeaux, France
Virginie.Garcia@bordeaux.inra.fr

Keywords Web-services, service-oriented architecture (SOA), RNA-Seq.

BIOS: une architecture orientée service pour l'analyse de données de RNA-seq

http://bios.toulouse.inra.fr

Mots-clés Web-services, architecture orientée services (SOA), RNA-Seq

We present BIOS, a Service-Oriented Architecture (SOA) for RNA-seq analysis. Through a unified web interface, users build and parameterize their analysis workflow, accessing in a transparent way the data and/or the analytic services proposed by a network of eight servers distributed in eight laboratories. The BIOS network gives access to data of several species of agronomic interest (plants, insects, oomycetes, etc.) as well as permits the identification of differentially expressed transcripts based on data provided by the user in a very simple tabulated format. Five flash tutorials illustrate the proposed analysis programs which are adapted to the various technologies (Sanger, 454, Illumina) used for measurements of expression based on sequence counts with [1] or without replicates [2,3,4]. The data and the analytic services are distributed; the communication between the application and the servers is performed by BioMoby [5] web-services registered in the BIOS central registry (ten web-services for data access, ten analytic web-services, one web-service for network management). In addition, BIOS web-services ensure the interoperability with external systems, allowing for example the integration of expression patterns from “gene report” applications. In order to guarantee a crucial and stable quality of service, the entire network is supervised, both at the hardware and software levels. Thus, functional tests of the web-services are carried out daily. The result of this monitoring is placed at the users' disposal in order to ensure the best possible quality of service and to provide a maximum of transparency.

The service-oriented architecture BIOS, applied to the RNA-seq problem, offers a great flexibility and scalability. Indeed, after being uploaded on one of the servers, data benefits immediately from all the analysis programs available on the network. Conversely, once a new program has been added on a node of the network it can immediately be used to analyze any data.

BIOS is currently used for the data analysis of eighteen species. To date, two publications citing BIOS have been released (*Biomphalaria glabrata* [6], 454 and illumina data and *Rosa chinensis* [7], Illumina data) and several others are in preparation. This website is free and open to all users and there is no login requirement (login gives access to unpublished data).

Acknowledgements

This work was supported by INRA (AOBD 2009-2010)

References

[1] Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome biology*, 11, R106

[2] Stekel,D.J., Git,Y. and Falciani,F. (2000) The comparison of gene expression from multiple cDNA libraries. *Genome research*, 10, 2055-61.

[3] Herbert,J.M.J., Stekel,D., Sanderson,S., Heath,V.L. and Bicknell,R. (2008) A novel method of differential gene expression analysis using multiple cDNA libraries applied to the identification of tumour endothelial genes. *BMC genomics*, 9, 153

[4] Susko,E. and Roger,A.J. (2004) Estimating and comparing the rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys. *Bioinformatics* (Oxford, England), 20, 2279-87

[5] Wilkinson,M.D., Senger,M., Kawas,E., Bruskiewich,R., Gouzy,J., Noirot,C., Bardou,P., Ng,A., Haase,D., Saiz,E. de A., et al. (2008) Interoperability with Moby 1.0--it’s better than sharing your toothbrush! *Briefings in bioinformatics*, 9, 220-31

[6] Deleury E, Dubreuil G, Elangovan N, Wajnberg E, Reichhart JM, Gourbal B, Duval D, Baron OL, Gouzy J, Coustau C. (2012) Specific versus non-specific immune responses in an invertebrate species evidenced by a comparative de novo sequencing study. PLoS One. , 7(3)

[7] Annick Dubois, Sebastien Carrere, Olivier Raymond, Benjamin Pouvreau, Ludovic Cottret, Aymeric Roccia, Jean-Paul Onesto, Soulaiman Sakr, Rossitza Atanassova, Sylvie Baudino, Fabrice Foucher, Manuel Bris, Jérôme Gouzy, Mohammed Bendahmane (2012) Transcriptome database resource and gene expression atlas for the rose. *BMC Genomics*, 13, 638

EuGene-PP: Automatic and comprehensive annotation of prokaryotic genomes with oriented RNA-Seq

Erika SALLET[1,2], Emmanuel COURCELLE[1,2], Thomas FARAUT[3,4], Thomas SCHIEX[5] and Jérôme GOUZY[1,2]

[1] Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, INRA, Castanet-Tolosan, F-31326, France.

[2] Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594 CNRS, Castanet-Tolosan, F-31326, France.

[3] Laboratoire de Génétique Cellulaire, UMR 444, INRA Castanet-Tolosan, F-31326, France

[4] Laboratoire de Génétique Cellulaire, UMR 444, ENVT, Castanet-Tolosan, F-31326, France.

[5] Mathématiques et Informatique Appliquées Toulouse (MIAT), UR875 INRA, Castanet-Tolosan, F-31326, France.

Prenom.nom@toulouse.inra.fr

Keywords gene prediction, prokaryotes, oriented RNA-Seq.

With the new generation of sequencing (NGS) technologies, bacterial and archeal genome projects now combine deep genomic sequencing with a variety of transcriptome libraries (see [1] for example). If the main motivation for transcriptome sequencing is usually the quantification of gene expression, the transcribed sequences generated by deep sequencing can also contribute to prokaryotic genome annotation by the elucidation of gene structural features, including transcription start sites (TSSs), 5' and 3' untranslated regions (UTRs), and the identification of non-coding RNA (ncRNA) genes. In the recent sequencing of bacterial and archeal genomes, the annotation has still been done manually due to the lack of appropriate tools to integrate RNA-Seq data [2]. Indeed, most existing prokaryotic gene finders [3] or higher level bacterial annotation system [4] are based on genomic sequence analysis and do not take into account available expression data in the structural prediction. Expert annotation using RNA-Seq data has been recently facilitated by the use of integrated tools such as VESPA [5] or MicroScope [6] which allows to simultaneously visualize genomic, transcriptomic, proteomic or syntenic data, but the ultimate curation process still remains laborious.

There is therefore a clear need for automated prokaryotic genome annotation tools able to integrate the variety of informative data that can be produced by second generation sequencing (or other high-throughput analyses such as tiling arrays and proteomics). The development of such prokaryotic gene finders allowing the prediction of coding sequences (CDSs) but also TSSs and ncRNA genes, should provide improved transcript quantification (based on mapped read counts on predicted transcripts), facilitated identification of regulatory sequences upstream of mapped TSSs and thus, easier analysis of gene regulation.

Because of the higher complexity of eukaryotic gene structures and the usual availability of transcribed sequences (such as Expressed Sequence Tags or ESTs) many eukaryotic gene finders already have the ability to integrate experimental evidence in their gene prediction process [7-8]. Among them, EuGene has been recently extended to produce a comprehensive genome annotation of prokaryotic genomes using high-throuput evidence [9].

Here, we present EuGene-PP (EuGene-Prokaryote Pipeline), a fully automatic and generic bacterial annotation pipeline capable of producing a qualitatively enriched structural genome annotation by combining the usual intrinsic information provided by coding potential (computed from Interpolated Markov Models), Stop and Start codon analysis (using a dedicated RBS alignment tool) with other information. EuGene-PP also integrates similarities with known proteins and high throughput strand-specific RNA-Seq data. Since it is very easy to integrate more information in Eugene, EuGene-PP also automatically integrates high quality CDS predictions (produced using the bacterial gene finder Prodigal [3]) and ncRNA predictions (produced by tRNAscan-SE, RNAmmer and rfam-scan software) as additional sources of evidences. The EuGene-PP installation therefore requires the installation of all these dependencies, and also of the mapping software glint (Faraut T. and Courcelle E.; http://lipm-bioinfo.toulouse.inra.fr/download/glint/, unpublished)

The Perl annotation pipeline EuGene-PP encapsulates the C++ integrative prokaryotic annotation tool we recently derived from EuGene [9]. It is based on the same integration principles as the EuGene eukaryotic gene finder, which can easily be described as a conditional random field. The main advantage of EuGene-PP is that is has extremely simple fully automatic use. In the basic case, it just requires a directory with genomic sequences, another with RNASeq read information (in FASTQ, FASTA, BAM, BED or WIG format) to run and produce the structural annotation in GFF3 format with the corresponding transcriptome and proteome in FASTA format, and some basic statistics (mean gene length, percentage of gene with UTRs, GC content, …). The pipeline, based on Paraloop software (http://lipm-bioinfo.toulouse.inra.fr/paraloop), is able to parallelize some time-consuming steps according to the user local configuration (cluster, multiprocessor system). Unsurprisingly, on *S. meliloti*, EuGene-PP provides an automatic annotation which is highly similar to the annotation described in [9]. More tests are needed to evaluate its robustness on a range of bacterial genomes.

EuGene-PP directly produces an enriched structural annotation using just the genomic sequence and the stranded RNA-Seq data, while minimizing manual expert annotation. All training procedures required for gene finding are performed inside EuGene-PP and are invisible to the end user. The produced annotation contains previously unpredicted important gene structure features such as 5' and 3'UTRs, and therefore TSS, as well as non-coding RNA genes (including antisense RNAs). The EuGene-PP pipeline will be soon made available as open-source software.

Acknowledgements

This work was supported by the grant ANR-08-GENO-106 "SYMbiMICS". This research was done in the Laboratoire des Interactions Plantes-Microorganismes, part of the Laboratoire d'Excellence (LABEX) entitled TULIP (ANR-10-LABX-41).

References

[1] Weissenmayer, B. A., Prendergast, J. G. D., Lohan, A. J. and Loftus, B. J., Sequencing illustrates the transcriptional response of Legionella pneumophila during infection and identifies seventy novel small non-coding RNAs, *Plos One*, 6, e17570. 2011.

[2] Richardson, E. J. and Watson, M., The automatic annotation of bacterial genomes, *Brief Bioinform*., 14, 1-12. 2013.

[3] Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W. and Hauser, .J., Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, 11, 119. 2010.

[4] Aziz Pareja-Tobes, P., Manrique, M., Pareja-Tobes, E., Pareja, E. and Tobes, R., BG7: A new approach for bacterial genome annotation designed for next generation sequencing data, *PLoS One*, 7, e49239. 2012.

[5] Peterson, E. S., McCue, L. A., Schrimpe-Rutledge, A. C., et al. 2012, VESPA: software to facilitate genomic annotation of prokaryotic organisms through integration of proteomic and transcriptomic data, *BMC Genomics*, **13**, 131.

[6] Vallenet, D., Belda, E., Calteau, A., et al, MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data, *Nucleic Acids Res*., 41, D636-47. 2013.

[7] Foissac, S., Gouzy, J., Rombauts, S., Mathé, C., Amselem, J., Sterk, L., van de Peer, Y., Rouzé, P., Schiex, T. Genome Annotation in Plants and Fungi: EuGène as a Model Platform. *Current Bioinformatics*, Volume 3, Number 2, p. 87-97, 2008

[8] Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. 2004, AUGUSTUS: a web server for gene finding in eukaryotes, *Nucleic Acids Res*., 32, W309-12.

[9] Sallet, E., RouxB., Sauviac L., Jardinaud, F., Carrère, S., Faraut, T., de Carvalho-Niebel, F., Gouzy, J., Gamas , P., Capela, D., Bruand, Caude and Schiex, T. Next Generation Annotation of prokaryotic genomes with EuGene-P: application to *Sinorhizobium meliloti*. *DNA Research*, 2013.

Private selective sweeps identified from next-generation pool-sequencing reveal convergent pathways in two inbred *Schistosoma mansoni* strains.

Julie A. J. Clément[1], Eve Toulza[1], Mathieu Gautier[2], Hugues Parrinello[3], David Roquis[1], Jérôme Boissier[1], Anne Rognon[1], Hélène Moné[1], Gabriel Mouahid[1], Jérôme Buard[4], Guillaume Mitta[1] and Christoph Grunau[1]

[1] Laboratoire Ecologie et Evolution des Interactions, UMR5244, CNRS - Université de Perpignan, 52 avenue Paul Alduy, F-66860, Perpignan, France

{julie.clement, eve.toulza, david.roquis, boissier, rognon, mone, mouahid, mitta, christoph.grunau}@univ-perp.fr

[2] Centre de Biologie pour la Gestion des Populations, UMR1062 (INRA – IRD – Cirad – Montpellier SupAgro), Campus International de Baillarguet, F-34988 Montferrier-sur-Lez, France.

mathieu.gautier@supagro.inra.fr

[3] MGX Montpellier GenomiX, 141, rue de la Cardonille, F-34094 Montpellier, France.

hugues.parrinello@mgx.cnrs.fr

[4] Institut de Génétique Humaine, UPR1142 CNRS, 141, rue de la Cardonille, F-34396 Montpellier, France.

jerome.buard@igh.cnrs.fr

Keywords Host-parasite interactions, evolutionnary genomics, population genomics, pool-sequencing, selective sweeps.

Trematode flatworm of the genus *Schistosoma*, the causative agent of bilharziasis, is among the most prevalent parasite in humans, affecting more than 200 million people worldwide. In this study, we focused on two well-characterized strains of *Schistosoma mansoni*, to explore signatures of adaptation underlying phenotypic variation. Both strains have previously been shown to be highly inbred and to exhibit differences in life history traits, in particular in their compatibility with the intermediate snail host *Biomphalaria glabrata*. Thanks to new sequencing technologies, it is now possible to provide a detailed picture of genetic variation on a genome-wide scale allowing in turn the identification of genomic regions subjective to strong selective pressure. For the purpose of this study, we sequenced pools of DNA from individuals belonging to both strains using Illumina technology at an overall 20X genome-wide coverage.

In contrast to results obtained with microsatellite markers, a high number of SNPs was identified despite of inbreeding through several decades. As expected, the two strains present moreover high inter-population differentiation. In total, 708,898 SNPs were identified and roughly 2,000 copy number variations which containing mostly genes encoding hypothetical proteins. Based on a recently developed test, we further identified 272 footprints of selection (146 overlapping in the two strains). Interestingly, functional annotation of protein-coding genes in regions under selection revealed functions related to parasitic lifestyle (e.g. cell-cell adhesion or oxydo-reduction) but also genes encoding vaccine candidate proteins.

We provide here the first comprehensive report about genomic diversity in two *S. mansoni* strains showing that even under inbreeding, the parasite still presents genomic single nucleotide and structural polymorphisms. Despite high differentiation between strains, we identified in private i.e. non-overlapping selective sweeps common convergent pathways, notably proteolysis.

Acknowledgements

We would like to thank Alexandre Nore for technical support during in situ validation of SNPs and Simon Boitard for helpful assistance on the implementation of the pool-hmm model. This work received funding from the French National Agency for Research (ANR), project ANR-2010-BLAN-1720-01 (EPIGEVOL).

MGX – Montpellier GenomiX facility

RNA-Seq analysis

Leila BASTIANELLI[1], Vincent DEMOLOMBE[1], Emeric DUBOIS[1], Sabine NIDELET[1], Hugues PARRINELLO[1], Stéphanie RIALLE[1], Dany SEVERAC[1] and Laurent JOURNOT[1]

MGX - Montpellier Genomix, c/o Institut de Génomique Fonctionnelle ; 141 rue de la cardonille, 34094 Montpellier, Cedex 05, France

{leila.bastianelli, vincent.demolombe, emeric.dubois, sabine.nidelet, hugues.parrinello, stephanie.rialle, dany.severac, laurent.journot}@mgx.cnrs.fr

Abstract *BioCampus Montpellier hosts several technical platforms including MGX - Montpellier GenomiX, a microarray and next-gen sequencing facility (Illumina HiSeq2000). This platform is aimed at seamless integration of data production with various data analysis tools. The MGX team comprises 3 molecular biologists, 4 bioinformaticians and 1 manager. The facility is accessible to both academic and industry/biotech scientists. Next-generation sequencing has many applications for which the MGX team has developed knowledge and expertise ; here as an example we present our RNA-Seq analysis pipeline.*

Keywords Genomics, next generation sequencing, RNA-Seq, bioinformatic analysis.

MGX – Montpellier GenomiX facility

Analyse RNA-Seq

Résumé *BioCampus Montpellier héberge plusieurs plateformes techniques, dont MGX - Montpellier GenomiX, une plateforme de microarrays et séquençage nouvelle génération (Illumina HiSeq2000). L'équipe MGX compte 3 biologistes moléculaires, 4 bioinformaticiens et 1 responsable. Elle est accessible à des scientifiques aussi bien académiques qu'issus de l'industrie. Le séquençage haut débit a de nombreuses applications, pour lesquelles l'équipe MGX a acquis des connaissances et de l'expertise ; nous présentons ici à titre d'exemple notre pipeline d'analyse RNA-Seq.*

Mots-clés Genomique, séquençage nouvelle génération, RNA-Seq, analyse bioinformatique.

1 Introduction

1.1 La plateforme MGX

La plateforme de service en génomique de Montpellier (MGX – Montpellier GenomiX) est une plateforme labellisée par le réseau IBiSA et le Cancéropôle Grand Sud-Ouest. Elle propose des services en microarrays, séquençage nouvelle génération et bioinformatique pour les communautés scientifique et industrielle, dans tous les domaines du Vivant. Depuis octobre 2010, en reconnaissance de sa conformité au référentiel ISO 9001 : 2008, le plateau technique IGF/IGH est certifié pour ses activités de développement et de réalisation de prestations en génomique (microarray, séquençage et bioinformatique). La plateforme apporte son expertise pour :

- conseiller les utilisateurs dans la réalisation de leurs expériences et leur proposer les plans expérimentaux les mieux adaptés
- générer les données microarrays et séquençage nouvelle génération
- analyser les données générées sur la plateforme
- former les utilisateurs aux outils mis en place et les accompagner dans l'interprétation des données.

1.2 RNA-Seq

La plateforme propose de nombreuses applications et techniques de séquençage : ADN génomique (séquençage de novo ou reséquençage, exome, RAD-Seq...), sites de liaisons de facteurs de transcription ou modifications d'histones (ChIP-Seq), ...
Parmi elles, le RNA-Seq (séquençage de transcriptome) permet d'obtenir des informations sur les ARN présents dans une cellule à un temps donné. On peut ainsi étudier l'ARN messager si on s'intéresse à l'expression des gènes, mais aussi les small RNA, ainsi que des phénomenes plus particuliers (par exemple le GRO-seq, qui séquence les ARN en cours de transcription au moment de l'extraction). Les ARN sont extraits (et éventuellement sélectionnés), puis rétro-transcrits en cDNA (ADN complémentaire) et enfin séquencés sur HiSeq2000. Le RNA-Seq peut être "directionnel" : dans ce cas, au moment de la reverse-transcription, un protocole adapté permet de conserver l'orientation du transcrit et donc éventuellement de différencier certains gènes par la suite.

2 Pipeline d'analyse RNA-Seq

La plateforme MGX découpe ses analyses en 3 catégories successives, correspondant à des niveaux de complexité croissante.

2.1 "Niveau 1" : Bioinformatique de base

Le niveau 1 est indissociable du séquençage en lui-même : il s'agit de l'analyse d'images, du "base-calling" (identification de chaque base séquencée à partir des images) et du contrôle qualité via FastQC.
Ce niveau comprend également l'alignement des séquences sur le génome ou le transcriptome de référence de l'espèce étudiée ; il est réalisé avec le module eland_rna de CASAVA pour les expériences en single-read, et avec TopHat pour les expériences en paired-end. Les deux méthodes génèrent un transcriptome virtuel à partir d'un fichier d'annotations, alignent les reads sur ce transcriptome puis alignent les reads restants sur le génome. TopHat permet également de détecter de nouvelles jonctions d'épissage.

2.2 "Niveau 2" : Analyses statistiques

L'analyse de l'expression différentielle des gènes commence par une étape de comptages : le logiciel HTSeq-count compte le nombre de reads mappés sur chaque transcrit, qui reflète son niveau d'expression. Ces comptages tiennent compte de l'orientation des reads s'il s'agit de RNA-Seq directionnel.
Les analyses statistiques d'expression différentielle sont ensuite réalisées avec deux packages R : EdgeR et DESeq. Ces analyses incluent une première étape de normalisation des données, qui permet de corriger une partie des biais auxquels sont soumis les comptages de RNA-Seq (taille des librairies, différence d'expression des gènes, taille des transcriptomes), via 3 normalisations différentes : TMM, RLE, UpperQuartile. Dans le cas d'échantillons appariés, une analyse statistique est réalisée grâce à la méthode GLM implémentée dans EdgeR.

2.3 "Niveau 3" : Contextualisation

Une fois la liste de gènes différentiellement exprimés (DE) obtenue, une étape de contextualisation peut être réalisée. Celle-ci se base sur les données du consortium Gene Ontology (GO), elle n'est réalisable que si les annotations GO sont disponibles pour l'organisme étudié (19 espèces disponibles). On évalue si les différents termes GO sont statistiquement sur-représentés dans la liste de gènes d'intérêt. On peut ainsi mettre en évidence des voies métaboliques régulées dans les conditions étudiées.
Un autre type d'analyse proposé par la plateforme est la classification hiérarchique. Elle permet de regrouper les gènes et échantillons qui présentent des profils d'expression similaires. On peut ainsi découvrir de nouvelles sous-classes d'échantillons et distinguer des groupes de gènes corégulés donc impliqués dans une même fonction biologique.
Une fois les analyses terminées, la plateforme aide également les clients à la publication des résultats en rédigeant les matériels et méthodes, en réalisant des figures/tableaux pour leur article et en déposant leurs données dans une base de données publique telle que Gene Expression Omnibus (GEO) du NCBI.

Efficient strategy to find new genes involved in Parkinson's disease

Anne-Sophie COQUEL[1,2], Christel CONDROYER[1,2] , Aurélie HONORE, Alexis BRICE[1,2,3] and Suzanne LESAGE[1,2]

[1]INSERM, UMR-S975, 83 boulevard de l'hôpital, 75013, Paris, France
anne-sophie.coquel@inserm.fr

{christel.condroyer, aurélie.honore, alexis.brice, Suzanne.lesage}@upmc.fr

[2] Université Pierre et Marie Curie-Paris6, UMR_S975, centre de recherche de la moelle et du cerveau CRICM, Pitié-Salpêtrière, 75013, Paris, France
[3] AP-HP, Hôpital Pitié-Salpêtrière, Département de génétique et cytogénétique, 75013, Paris, France

Keywords Parkinson's disease, autosomal recessive, homozygosity mapping, whole exome.

Parkinson's disease (PD) is the second most prevalent neurodegenerative disease after Alzheimer's disease. It affects 1-2% of individuals over the age of 60. More than 100 000 patients suffer from this disease in France. PD typically manifests with impairments of motor function, rigidity, slowness of movement and poor postural stability, caused by the loss of dopaminergic neurons in the substantia nigra of the midbrain. So far treatments are only substitutive and do not prevent the massive neurodegeneration. The cause of PD is poorly understood but in literature there are evidences suggesting interplay between genetic and environmental factors. During the last decade, a total of 18 PD loci have been identified through linkage analysis (Park 1-15) or genome wide association studies (Park 16-18). Mutations in genes at 6 (SNCA, Parkin, PINK1, DJ- 1, LRRK2, ATP13A2) of these 18 loci have been conclusively associated with inherited forms of Parkinsonism. To date, the known mutations account for no more than 5% of all cases of PD and it can be estimated that approximately 50% of recessive and probably more than 80% of dominant familial cases are caused by as yet unknown mutations.

To identify novel genes for recessive PD, we use an efficient integrative approach combining homozygosity mapping/genomic rearrangement detection and targeted exome sequencing in consanguineous PD families with an early age at disease onset (≤55 years). Through the French network for the study of Parkinson's disease genetics and diverse collaborations with Mediterranean countries, we have already collected one of the largest series of PD families or isolated patients with self-reported known or suspected consanguinity or originating from the same location. Homozygosity mapping was performed in a subset of 160 PD consanguineous families or isolated patients excluded for PARK2, PINK1, DJ-1, and the common LRRK2 G2019S mutation, using single nucleotide polymorphism (SNP) genotyping microarrays (by using GenomeStudio software to define the homozygous areas for each patient) and an original linkage statistics program that allows the inclusion of individuals for whom genealogical information is lacking.
Computation of the inbreeding coefficients F using FEstim that was developed by our collaborator, A.L. Leutenegger (INSERM UMR_S946) allows detecting any individuals without consanguinity (F=0) and to remove them from further analyses. These inbreeding coefficients F along with multipoint genetic linkage analyses were computed for all consanguineous families and isolated cases with the Lod score statistic (FLOD) that incorporates the inbreeding coefficients F of each individual. These statistical analyses allow identifying inbred individuals and distinguishing regions that are autozygous from the regions that are only homozygous by chance. We prioritize segments of homozygosity shared by the largest number of consanguineous families. Besides, we discovered a homozygous area shared by 13 PD patients on chromosome 1 with a LOD score >3. We are investigating this area.

Using the same SNP genotyping microarrays, we search for rare copy number variations (CNVs) (large deletions and duplications). To identify causative genes, we sequenced the exome of a series of 70 rearly-onset consanguineous patients (16 families, 9 of which have at least two members done in whole exome sequencing; 45 isolated cases) using the combination of exome capture and next generation DNA sequencing. The annotation of the whole exome data has been done by implementing the Annovar pipeline

and home-made R scripts to select the candidate genes and variants. After filtering the large amount of variants identified in these patients against public databases such as dbSNP, 1000 Genomes and NHLBI exome sequencing projects and then stratifying candidate variations by their functional class (i.e frameshift, stop codons and splice sites versus missense variants) and/or by existing biological, functional, expression and conservation information about these candidate genes, we systematically search for genes with rare homozygous mutations or rearrangements (frequency of the heterozygous variants <1%). The
regions of homozygosity by descent and the homozygous genomic rearrangements detected, particularly when they are shared by several families, help to refine the candidate regions for analysis of the exome data.
Without taking into account the variant frequency, each patient has between 1500 and 2000 homozygous variants in homozygous areas and after filtering for the effect (non-synonymous, stop codon, frameshit, splice site), each patient has between 1 and 30 homozygous variants.

This efficient strategy allows us identifying 2 genes in one consanguineous family. The variants are heterozygous for the parents and the healthy brother and homozygous for the PD child. One gene is more interesting and hasn't been found in our cohort of 532 controls (whole exome).
4 other candidate genes have been identified in isolated consanguinous cases.
All the variants have been checked for each family or isolated cases by Sanger sequencing. At the present time, we are looking for other variants in these genes in a cohort of 96 AR PD patients and 96 controls. As a test of our strategy, by looking for genes and variants in homozygous areas, we found again the mutation in the ATP13A2 gene in a family with PD patients.

The identification of new genes involved in early-onset autosomal recessive PD will not only contribute to diagnosis and genetic counseling in patients and their relatives, but will also open new avenues of research into the mechanism of neurodegeneration in PD, and the development of innovative scientifically based treatments to cure or slow progression of the disease, which do not yet exist.

Evolution of Repeated Extragenic Palindromic Sequences

Application in *Escherichia* and *Shigella* genomes

Mathias WEYDER[12], Patricia SIGUIER[12], Bao TON-HOANG[12], Mick CHANDLER[12], Gwennaele FICHANT[12] and Yves QUENTIN[12]

[1] Laboratoire de Microbiologie et Génétique Moléculaires, UMR5100 CNRS, 118 Route de Narbonne, 31062, Toulouse, Cedex, France

(weyder, siguier, tonhoang, chandler, fichant, quentin)@ibcg.biotoul.fr

[2] Université de Toulouse, UPS, Laboratoire de Microbiologie et Génétique Moléculaires, Toulouse, France.

Keywords Repeat elements, annotation, secondary structure, evolution, genome plasticity

1 Contexte et objectifs

Initialement découvertes chez les entérobactéries, notamment chez *Escherichia coli K12* où elles composent environ 1% du génome [1], les séquences REP (Repeated Extragenic Palindromes) sont très largement répandues dans le règne bactérien [2]. Les REP sont majoritairement retrouvées sous forme de structures extragéniques fortement répétées appelées BIMEs (Bacterial Interpersed Mosaic Element). Une BIME est composée de deux REP en orientation inversées, reliées par une séquence variable. Chez *E. coli*, nous distinguons trois classes de REP (repY, repZ1 et repZ2) qui se différencient aussi bien par leurs séquences consensus que par leurs longueurs. De par leur séquence primaire conservée mais aussi leur structure secondaire formant une tige boucle, de nombreuses propriétés sont associées aux REP (terminatrices de transcription, stabilisatrices d'ARNm). Leur capacité d'interaction avec plusieurs facteurs protéiques dont notamment l'ADN pol I, l'ADN gyrase et le facteur IHF a également été démontrée. Cependant, les fonctions qui découlent de ces propriétés demeurent toujours assez mal connues.

Il a été proposé que la protéine TnpA_{REP}, qui ressemble aux transposases de la famille Y1, soit responsable de la prolifération des REP. Le gène codant pour la TnpA_{REP} est encadré par un nombre variable de BIME [3]. Ce gène n'est pas présent dans toutes les souches de *E. coli* et quand il est présent il n'est pas toujours fonctionnel. Par une approche *in vitro,* il a démontré, que seules les REP possédant une conformation en tige-boucle, le tétranucléotide GTAG conservé à la base de la structure, ainsi qu'un mésappariement dans la région intermédiaire de la tige, avaient la capacité d'une part d'interagir avec la TnpA_{REP} [4] et d'autre part d'être clivées et recombinées par la TnpA_{REP} [3].

Nous proposons de tester la prolifération des REPs au travers d'une approche de génomique comparative. Pour cela, nous devons 1) annoter les gènes codant pour la TnpA_{REP}, 2) annoter les REPs dans les différents génomes d'*Escherichia coli* et de *Shigella* disponibles, 3) identifier les REP orthologues et 4) replacer les évènements de gains/pertes de REP sur l'arbre phylogénétique des souches/espèces analysées.

2 Méthodes et résultats

Les génomes complets des souches d'*Escherichia* et *Shigella* ont été obtenus à partir du site du NCBI. L'annotation du gène codant pour la TnpA_{REP} peut être réalisée par des approches classiques (BlastN). Par contre, l'annotation des REP est plus délicate en raison de leurs petites tailles (entre 20 et 40 nucléotides), de leur hétérogénéité de conservation, de leur arrangement en tandem et également à leur propension à servir de sites d'intégration d'IS. Une autre difficulté est de distinguer entre une vraie REP et une séquence ressemblant à une REP. Nous avons utilisé les REP annotées manuellement chez *E. coli K12* par Sophie Bachelier pour évaluer les performances des différentes approches. Nous avons donc choisi de prédire les REP individuellement et, dans une seconde étape, d'assembler les REP associées en tandem sous forme de BIME.

Nous avons envisagé trois stratégies, la première repose sur un alignement local (BlastN) entre les séquences consensus des trois familles de REP de *E. coli* et les séquences génomiques. La seconde, basée sur GLAM [5], utilise des profils calculés sur les membres des familles de REP. La troisième utilise le modèle de covariance du logiciel Infernal [6]. Ces trois méthodes ont des propriétés différentes. Les paramètres de chacune des méthodes sont étalonnés par rapport à l'annotation manuelle des REP effectuée par S. Bachelier. La méthode basée sur un alignement local présente l'inconvénient de ne pas toujours prédire correctement les bornes des REP. Elle est rapide mais peu sensible. Elle est donc écartée de la prédiction des REP mais nous la conservons dans notre chaine de traitement car elle permet de prédire les IS et le gène *tnpA*$_{\mathrm{REP}}$. La méthode basée sur GLAM est un peu moins performante que celle basée sur Infernal. La prise en compte de la structure secondaire augmente significativement la sensibilité de la prédiction. Infernal est ainsi utilisé pour effectuer les prédictions. Les REP prédites sont regroupées en une même BIME si elles sont séparées par moins de 70 bp, valeur apprise sur l'échantillon de S. Bachelier.

Nous avons appliqué notre chaine de traitement à 71 génomes complets d'*Escherichia* et *Shigella*. Les annotations des fichiers GenBank et les résultats des annotations sont entrés dans une base de données spécialisée. Les caractéristiques de séquence primaire et secondaire des REP, pouvant être déterminant pour leur mobilisation par la TnpA$_{\mathrm{REP,}}$ ont ensuite été recherchées. Pour l'évaluation des structures secondaire nous avons utilisé le logiciel MFOLD [7]. Pour identifier les REP orthologues, nous nous basons sur l'alignement des génomes obtenus par MAUVE [8]. Cet alignement nous permet d'identifier les gènes orthologues bornant les *loci* contenant des REP, ainsi considérées comme orthologues.

Les résultats obtenus montrent que si l'on considère le flux global de REPs le long de l'arbre phylogénétique des souches étudiées, l'augmentation du nombre moyen de REP par génome le plus important est observé pour les souches du phylogroupe A, possédant toutes le gène *tnpA*$_{REP}$. Néanmoins, si nous considérons chaque *locus* orthologue indépendamment, il semble exister des événements de variations de tailles de BIME en absence de ce gène, suggérant qu'il existerait d'autres mécanismes conduisant à une fluctuation du nombre de REP associées en BIME. Par ailleurs, l'acquisition de ce gène par transfert horizontal dans des souches où le gène avait été perdu chez l'ancêtre, n'a pas conduit à une augmentation du nombre de REP.

Remerciements

This work was supported by ANR-12-BSV8-0009-02.

Références

[1] S. Bachellier, W. Saurin, D. Perrin, M. Hofnung and E. Gilson, Structural and functional diversity among bacterial interspersed mosaic elements (BIMEs). *Mol. Microbiol,* 12:61–70, 1994.

[2] R. Tobes, E. Pareja, REP code: defining bacterial identity in extragenic space. *Environ Microbiol, 7:225–228, 2005.*

[3] B. Ton-Hoang, P. Siguier, Y. Quentin, S. Onillon, G .Fichant and M. Chandler, Structuring the bacterial genome, Y1-transposases associated with REP-BIME sequences. *Nucleic Acids Re*s, 40:3596–3609, 2012.

[4] SAJ. Messing, B. Ton-Hoang, AB. Hickman, AJ. McCubbin, GF. Peaslee, R. Ghirlando, M. Chandler and F. Dyda, The processing of repetitive extragenic palindromes: the structure of a repetitive extragenic palindrome bound to its associated nuclease. *Nucleic Acids Re*s, 40:9964–9979, 2012.

[5] TL. Bailey, M. Boden, FA. Buske, M. Frith, CE. Grant, L. Clementi, J. Ren, WW. Li, WS. Noble, MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, 37:W202–208, 2009.

[6] SR. Eddy and R. Durbin, RNA sequence analysis using covariance models. *Nucleic Acids Re*s, 22:2079–2088, 1994.

[7] M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res,* 31:3406–3415, 2003.

[8] ACE. Darling, B. Mau, FR. Blattner, NT. Perna, Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Res*, 14:1394–1403, 2004.

FlexFlux

A Java library to perform flux balance analyses with flexible regulatory constraints

Lucas MARMIESSE[1] and Ludovic COTTRET[1]

INRA, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, F-31326 Castanet-Tolosan, France
lucas.marmiesse@toulouse.inra.fr
ludovic.cottret@toulouse.inra.fr

Keywords metabolism, modelling, flux balance analysis, optimisation

1 Introduction

Flux Balance Analysis (FBA) is a widely used modelling approach for analysing genome scale metabolic networks [6]. To circumvent the lack of kinetic parameters for all the enzymes identified in an organism, this approach is based on constraints applied to the whole set of fluxes potentially active in a metabolic network. Two kinds of constraints are always present, whatever the method that uses FBA:

– the steady state constraint ensuring that each amount of metabolite being produced is consumed
– a lower and an upper bound for each reaction flux

The next step of FBA is to define an objective function, i.e a reaction or a set of reactions whose flux will be minimized or maximized. Classically, the objective function is a biomass reaction that contains all the metabolites that participate to the cell growth (lipids, amino acids, cofactors...) but one can choose to maximize a reaction producing a metabolite of biotechnological interest. By tuning flux bounds, it is possible to mimic environmental conditions (by allowing only some input fluxes) or genetic conditions (by turning off reactions linked to some genes) and to compute an optimal growth rate or optimal interesting metabolite production rate in these conditions. The problem is solved by using wel known linear programming algorithms. More than 100 different methods have been inspired by FBA to deal with problems that FBA alone can not resolve [5]. Some of them have been implemented in librairies such as Cobra-Toolbox [7] or Surrey-FBA [3]. However, it remains difficult for a modeller to create a new FBA-based method with the code supplied in these libraries. Moreover, it is still difficult for the modeller (biologist or bioinformatician) to add complex constraints to an existing model. To circumvent these drawbacks, we are implementing FlexFlux, the first JAVA library that performs FBA and that allows to easily specify complex regulatory constraints.

2 FlexFlux philosophy

We mainly focus the development of FlexFlux on two features : flexibility and speed. The flexibility of FlexFlux is at two levels:

– Flexibility for the development. The library is highly modular and new FBA methods are thus easy to develop. Moreover, several optimisations solvers can be bound to FlexFlux without many efforts.
– Flexibility for the end-user. The regulatory constraints and the condition values are set in files easy to fill in. New biological questions are thus easy to set in FlexFlux.

The speed of FlexFlux is ensured by the fact that each optimisation run can be parallelized to ensure velocity in analyses such as Flux Variability Analysis (FVA) [4] that perform successive FBA. Furthermore, as in the fastFVA algorithm [4], since the optimisation problem does not change a lot between two iterations of such a method, we use the warm start functions of the optimisation libraries that use the result of the previous iteration to make faster the resolution of the new iteration.

3 FlexFlux architecture

FlexFlux is based on the parseBioNet JAVA library that allows to load, edit and analyse metabolic networks (not published). The starting point of FlexFlux is the BioNetwork class of parseBioNet (**??**). A BioNetwork instance contains all the reactions, metabolites, genes and proteins that are involved in a metabolic network. Especially, it contains all the links between genes proteins and reactions (GPR) and the lower and upper bounds for each reaction flux. A BioNetwork instance can be created from a SBML file (exchange format for metabolic networks) or from a tabulated file that lists the reactions and their features. From a BioNetwork instance and an objective function, FlexFlux can already perform classical FBA based analyses such as Flux Variability Analysis, knock out simulations, etc... Additional constraints (genetic or environmental) that will overwrite the original constraints contained in the BioNetwork instance can be specified in the constraint file. In the same file, the modeller can also declare variables that don't exist in the initial BioNetwork. They are either numerical, integer or boolean and can correspond to transcription factors, abiotic conditions, life stage, etc... Regulatory rules are specified in an interaction file. Each interaction involves the variables specified in the BioNetwork instance (reactions, metabolites, etc...) and the ones specified in the constraint file. Regulatory constraints are coded in FlexFlux thanks to the relation "IF .. THEN". For instance, the rule "IF(TF1 OR TF2) THEN (R1 $>$ 2)" means that the presence of a transcription factor TF1 or the presence of the transcription factor TF2 makes that the flux towards the reaction R1 is greater than 2. To say that the inverse "R1 $>$ 2 implies TF1 or F2" is also true, the modeller can write "(TF1 OR TF2) EQ (R1 $>$ 2)", what is equivalent to two "IF ... THEN" relations.

The constraints and the logical relations are stored in specific FlexFlux objects. If integer variables have been declared or if there are some logical relations, the problem is translated into mixed integer programming (MIP) statements. Otherwise, the problem is translated into linear programming (LP) statements. Then, an external optimisation library solves the problem. For the moment, only CPLEX is bound to FlexFlux but we plan to develop binds to other optimisation library (e.g GLPK).

4 FlexFlux end-user functions

Currently, the methods implemented in FlexFlux are: FBA, fastFVA, gene/reaction knock-out simulations. In the next months, we are going to develop methods that take into account regulatory information: R-FBA [2], SR-FBA [8] and PROM [1].

References

[1] Sriram Chandrasekaran and Nathan D Price. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(41):17845–50, October 2010.

[2] M W Covert, C H Schilling, and B Palsson. Regulation of gene expression in flux balance models of metabolism. *Journal of theoretical biology*, 213(1):73–88, November 2001.

[3] Albert Gevorgyan, Michael E Bushell, Claudio Avignone-Rossa, and Andrzej M Kierzek. SurreyFBA: a command line tool and graphics user interface for constraint-based modeling of genome-scale metabolic reaction networks. *Bioinformatics (Oxford, England)*, 27(3):433–4, February 2011.

[4] Steinn Gudmundsson and Ines Thiele. Computationally efficient flux variability analysis. *BMC bioinformatics*, 11(1):489, January 2010.

[5] Nathan E Lewis, Harish Nagarajan, and Bernhard ØPalsson. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature reviews. Microbiology*, 10(4):291–305, April 2012.

[6] Jeffrey D Orth, Ines Thiele, and Bernhard ØPalsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–8, March 2010.

[7] Jan Schellenberger, Richard Que, Ronan M T Fleming, Ines Thiele, Jeffrey D Orth, Adam M Feist, Daniel C Zielinski, Aarash Bordbar, Nathan E Lewis, Sorena Rahmanian, Joseph Kang, Daniel R Hyduke, and Bernhard ØPalsson. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc*, 6(9):1290–1307, September 2011.

[8] Tomer Shlomi, Yariv Eisenberg, Roded Sharan, and Eytan Ruppin. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular systems biology*, 3:101, January 2007.

Orphan enzyme survey, a decade later

Maria Sorokina[1,2,3] , Mark Stam[1,2,3], Karine Bastard[1,2,3] , Claudine Medigue[1,2,3], Olivier Lespinet[4] and David Vallenet[1,2,3]

[1] CEA, IG, Genoscope,2 rue Gaston Crémieux CP5702, F-91057, Évry, France

[2] CNRS, UMR 8030, 2 rue Gaston Crémieux, CP5702, F-91057, Évry, France

[3] Université d'Evry Val d'Essonne, F-91057, Évry, France

[4] Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris Sud, Bâtiment 400, 91405 Orsay Cedex, France

msorokina@genoscope.cns.Fr

Keywords Orphan enzymes, metabolism

Millions of protein database entries are not assigned reliable functions. This shortcoming limits the knowledge that can be extracted from genomes. In contrast, the «orphan enzyme activities» problem, which was reported for the first time a decade ago[1, 2], corresponds to experimentally characterized activities that lack associated protein sequence. Here, we present an update of previously conducted surveys on orphan enzymes[3–5].

While the percentage of orphan enzymes has decreased from 38% to 22% in ten years, they are still more than 1,000 among the 5,000 entries of the Enzyme Commission classification. Though, the number of EC entries has increased considerably in the last years: more than 800 were created since 2010. We extended this study to local orphans: activities which have no representative sequence in a given clade but have one in other organisms. We observed an important bias in Archaea for local orphans and estimated the presence of candidate homologous proteins that may be shared between Eukaryotes and Prokaryotes. Beside, an analysis of orphan enzyme connectivity in metabolic networks was made: it shown that many of them are not in a pathway and only few ones are neighbors of non-orphan activities. Finally, by studying relations between protein domains and catalyzed activities, we showed that newly discovered enzymes are mostly associated with already known enzyme domains. Thus, the exploration of the promiscuity of known enzyme families may solve a part of the orphan enzymes. Indeed, the discovery of new families may extend the landscape of enzymatic activities.

References

[1] Karp, P. D. Call for an enzyme genomics initiative. *Genome biology* **5**, 401 (2004)

[2] Lespinet, O. & Labedan, B. Orphan enzymes? *Science (New York, N.Y.)* **307**, 42 (2005).

[3] Chen, L. & Vitkup, D. Distribution of orphan metabolic activities. *Trends in biotechnology* **25**, 343–8 (2007).

[4] Pouliot, Y. & Karp, P. D. A survey of orphan enzyme activities. *BMC bioinformatics* **8**, 244 (2007).

[5] Hanson, A. D., Pribat, A., Waller, J. C. & De Crécy-Lagard, V. “Unknown” proteins and “orphan” enzymes: the missing half of the engineering parts list--and how to find it. *The Biochemical journal* **425**, 1–11 (2010).

VENOMICS: High-throughput transcriptomics annotation of animal venom cell-glands for discovery of novel venom peptides.

http://www.venomics.eu

Marion VERDENAUD[1], Sheila ZUNIGA[2], Juan Carlos TRIVINO[2], Edwin de PAUW[3], Loic QUINTON[3], Michel Degueldre[3], Pierre ESCOUBAS[4] and Frédéric DUCANCEL[1]

[1] iBiTEc-S/SPI Antibody Engineering for Health Laboratory (LIAS), CEA, 91191 Gif-sur-Yvette, France
{marion.verdenaud, frederic.ducancel}@cea.fr

[2] Departments of New Technologies, R+D+I and Bioinformatics, Sistemas Genomicos Ltd, Valencia, Spain
{sheila.zuniga, jc.trivino}@sistemasgenomicos.com

[3] Laboratory of Mass Spectrometry, Department of Chemistry, University of Liege, Liege, Belgium
{e.depauw, loic.quinton, mdegueldre}@ulg.ac.be

[4] Venometech, 473 Route des Dolines – Villa 3, 06560 Valbonne, France
escoubas@venomethech.com

Keywords RNASeq, Toxins, Annotations, Venom compounds.

Venomous animals have developed an arsenal of small-reticulated compounds (mainly peptides) used in defense and predation. Based on various disulfide-linked scaffolds, they represent an enormous structural and pharmacological diversity. Molecular structures and the pharmacological spectrum of venom peptides are very diverse, making them top candidates as innovative drug leads [1,2]. Mass spectrometry and transcriptomics studies have shown the presence of several hundreds of peptides and proteins in the venom of single species of cone snails and spiders. Therefore the global animal venom resource can be seen as a collection of more than 40,000,000 peptides and proteins of which only ~3000 are known.

Nevertheless the use of venoms for drug discovery is a rapidly emerging but still mostly unrealized prospective, due to several major difficulties including the availability of material, the sample size (most venomous animals are small to very small) and the complexity of venoms. Instead of relying on the classical, low-throughput bioassay-guided approach to bioactive peptide identification, the European project VENOMICS proposes a totally new paradigm that completely bypasses the classical approach by *combining transcriptomics* and *proteomics* technologies to access venom diversity through peptide sequences instead of purification [3,4,5], followed by *in vitro* production of peptides [6], for use in drug screening programs.

One of the key-point of this project is the capacity to predict the exact sequence of the mature active peptides synthesized by the venom gland cell machinery. To reach that goal, the exact N-ter extremities of the toxins and venom compounds has to be determined, and also the possibility of most post-translational modifications identified (PTMs).

One the one hand, proteomics analysis is carried out on the venom itself that is milked from venomous animals. The venom is partitioned using HPLC in approximately 30 fractions. For each fraction, the protein disulfide bridges are reduced and purified before masses analyses using MALDI-In Source Decay TOF MS [7,8]. Sequences tags from each fraction are obtained using innovative fragmentation method as MALDI-ISD [9].

On the other hand, mRNAs are extracted from venom gland cells and sequenced by Illumina/Solexa platform. The reads are qualified, cleaned with in-house scripts and then are assembled into contigs using Oases with several k-mer values and CAP3 [10,11,12]. To predict the open reading frames (ORFs), we use framedp software [13]. The general annotation process is done by combining Blast and InterProScan annotations [14,15,16]. Special attention and efforts have been made for the "toxin-like" annotation and the mature sequence prediction. We use in-house scripts, known tools and curated databases to predict signal, propeptide and mature sequences [17,18,19].

Finally, the results of these two strategies are combined to determine the exact sequence of the mature peptide and the possible post-translational modifications. The new identified toxins are meant to be synthesized using chemical or recombinant ways, then tested towards different biological targets to identify potential new therapeutic hits.

Acknowledgements

VENOMICS is a European project supported through the Seventh Framework Program (FP7 HEALTH) 2011-2015.

References

[1] RJ. Lewis, ML. Garcia. Therapeutic potential of venom peptides. *Nat Rev Drug Discov*. 2(10): 790-802, 2003.

[2] Vetter, I., et al., Venomics: a new paradigm for natural products-based drug discovery. *Amino acids,* 40(1): 15-28, 2010.

[3] P. Escoubas, GF King. Venomics as a drug discovery platform. *Expert Rev Proteomics.* 6(3): 221-224, 2009.

[4] Wagstaff et al, Combined snake venomics and venom gland transcriptomic analysis of the ocellated carpet viper, Echis ocellatus. *J. Proteomics* 71, 609-623, 2009.

[5] Escoubas P, Sollod B, King GF. Venom landscapes: mining the complexity of spider venoms via a combined cDNA and mass spectrometric approach. *Toxicon: official journal of the International Society on Toxinology* 47(6): 650-63.,2006.

[6] Gräslund S et al, Protein production and purification. *Nat Methods* 5(2): 135-146, 2008.

[7] Quinton L, Demeure K, Dobson R, Gilles N, Gabelica V, De Pauw E. New methodfor characterizing highly disulfide-bridged peptides in complex mixtures : application to toxin identification from crude venoms. *J Proteome Res*. 2007 Aug;6(8):3216-23

[8] Ueberheide and al, Rapid sensitive analysis of cysteine rich peptide venom components. *Proc Natl Acad Sci USA* 106(17): 6910-15, 2009.

[9] Debois D, Smargiasso N, Demeure K, Asakawa D, Zimmerman TA, Quinton L, De Pauw E. MALDI in-source decay, from sequencing to imaging. *Top Curr Chem*. 2013;331:117-41

[10] Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28,1086-1092, 2012.

[11] Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* 29, 644-652, 2011.

[12] Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Research* 9, 868-877, 1999.

[13] Gouzy J, Carrere S, Schiex T. FrameDP: sensitive peptide detection on noisy matured sequences. *Bioinformatics*. 25(5): 670-1, 2009.

[14] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009 Dec 15;10:421

[15] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. InterProScan: protein domains identifier. *Nucleic Acids Res.* Jul;33. 2005.

[16] Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005 Sep 15;21(18):3674-6

[17] Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*. 2004 Jul 16;340(4):783-95.

[18] Jungo F, Bougueleret L, Xenarios I, Poux S. The UniProtKB/Swiss-Prot Tox-Prot program: A central hub of integrated venom protein data. *Toxicon.* 2012 Sep15;60(4):551-7.

[19] Kaas Q, Yu R, Jin AH, Dutertre S, Craik DJ. ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res*. 2012 Jan;40(Database issue)

EMBRC-France e-infrastructure

Marine Model Organisms Database

Gwendoline ANDRES[1], Alexandre CORMIER[2], Mylène LORRE-GUIDT[4], Mark HOEBEKE[1], Michel GROC[4], Philippe DRU[3] and Christophe CARON[1]

[1] STATION BIOLOGIQUE DE ROSCOFF, FR2424 CNRS-UPMC, ABiMS, 29680, Roscoff, France
{christophe.caron, gwendoline.andres, mark.hoebeke}@sb-roscoff.fr

[2] STATION BIOLOGIQUE DE ROSCOFF, UMR7139, Algal Genetics Group, CNRS-UPMC,
29680, Roscoff, France
alexandre.cormier@sb-roscoff.fr

[3] LABORATOIRE DE BIOLOGIE DU DEVELOPPEMENT, UMR7009 CNRS-UPMC, Observatoire Océanologique,
BP 28, 06234 Villefranche sur mer, France
philippe.dru@obs-vlfr.fr

[4] OBSERVATOIRE OCEANLOGIQUE DE BANYULS, UMS2348 CNRS-UPMC,
18, Av. du Fontaulé, 66650 Banyuls sur mer, France
{mylene.lorre-guidt, michel.groc}@obs-banyuls.fr

Keywords marine genome model, database, visualization, galaxy, marine biology

1 Introduction

The national infrastructure EMBRC-France will build a shared infrastructure based on the 3 UPMC/CNRS marine stations (Banyuls, Roscoff, Villefranche sur mer) to provide resources access. To promote exploitation of the resources, we need to integrate data and tools, and to provide workflow analysis capability. For the main terrestrial model organisms, international collaborative efforts have standardized and structured access to biological data and associated information through creation of open-access web resources providing access to a broad range of information about the given model organism. These centralized repositories MGI (Mouse), FlyBase (Drosophila), XenBase (Xenopus) and TAIR (Arabidopsis) are heavily exploited by research communities working on these model species and help to structure communities and promote use of biological and genetic resources.

2 E-infrastructure

An important component of this marine infrastructure is the creation of organism-focused web environment for flagship marine model organisms providing focal points for wider research communities. This dedicated database environment will be developed for one of the flagship marine model organisms in each of the main organism categories covered by EMBRC-France: prokaryotes (*Vibrio*, etc.), macroalgae (*Ectocarpus siliculosus*, etc.), microalgae (*Ostreococcus tauri*, etc.), and metazoans (*Clytia hemisphaerica*, etc.). In each case specific features of the organism, the status of currently available data and the requirements of the user community will be taken into account in developing both the database and the integrated specific analytical tools like specific collections of samples, to expression profile image data, or to access to specific mutants available in the concerning species.

This e-infrastructure is built by using standard components: Chado [1], Jbrowse, BioMart, BioDAS, etc. These frameworks (*e.g.* BioMart) facilitate interoperability with external resources and allow cross-query data from multiple database resources.

First, a web 2.0 portal has been designed to allow a standard unique and interactive access to resources databases and cross-relating –omics/imaging data, served by adapted visualisation user interfaces. The PostgreSQL environment is used with Chado data model, in order to provide a reliable tool. Finally, Galaxy [3] is deployed for on-line high throughput data analysis and to allow sharing data and analysis results.

3 Conclusion & Perspectives

EMBRC-France e-infrastructure provide a completely interoperable framework relevant for the French marine stations involved in EMBRC-France project, but also for widespread initiatives such as the ESFRI projects: EMBRC (http://www.embrc.eu/) and ELIXIR (http:// http://www.elixir-europe.org/). It will thus deliver a comprehensive suite of bioinformatics data and tools delivered to the community in a single environment. This will contribute significantly to national and international coordination and interoperability of the EMBRC e-infrastructure.

Acknowledgements

This work was supported by ANR program ‘Investissements d’Avenir’ (EMBRC-France), and the European Strategy Forum on Research Infrastructures Roadmap (EMBRC)

References

[1] A Chado case study: an ontology-based modular schema for representing genome-associated biological informationa comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.

[2] B. Giardine, C.Riemer, R.C.Hardison, R.Burhans, L.Elnitski, P.Shah, Y.Zhang, D.Blankenberg, I.Albert, J.Taylor, W.Miller, W.J.Kent and A.Nekrutenko, Galaxy: A platform for interactive large-scale genome analysis. *Nucleic Acids Res.*, 15: 1451-1455, 2005.

Identification of female sex locus in the brown alga *Ectocarpus*

Reference genome and *de novo* approaches to identify sex specific female scaffolds and female genes.

Alexandre CORMIER[1], John BOTHWELL[2], Mark J. COCK[1], Erwan CORRE[3], and Susana COELHO[1]

[1] STATION BIOLOGIQUE DE ROSCOFF, UMR7139, Algal Genetics Group, CNRS-UPMC, Place Georges Teissier, 29688, Roscoff, CS 90074, France

{acormier, cock, coelho}@sb-roscoff.fr

[2] DURHAM UNIVERSITY, Department of Biological Sciences, Stockton Rd, Durham, Durham City, County Durham DH1 3LE, UK

j.h.bothwell@durham.ac.uk

[3] STATION BIOLOGIQUE DE ROSCOFF, Plateforme ABiMS, FR2424, CNRS-UPMC, Place Georges Teissier, 29688, Roscoff, CS 90074, France

corre@sb-roscoff.fr

Keywords RNA-seq, *de novo* and transcriptome assembly, genome assembly, Sex-determining Region (SDR)

1 Introduction

Les bases moléculaires du déterminisme génétique du sexe et de la différenciation entre mâles et femelles ont beaucoup été étudiées chez les mammifères, les plantes et les champignons, mais rien n'est connu jusqu'à maintenant sur les mécanismes de détermination du sexe chez d'autres eucaryotes comme les algues brunes. L'identification et la caractérisation du locus contrôlant le caractère sexuel chez l'algue brune modèle *Ectocarpus* (dont le génome d'un individu mâle a été entièrement séquencé[1]) sont en cours dans l'équipe Génétique des Algues de l'UMR 7139 de la Station Biologique de Roscoff.

Nous avons utilisé des techniques de séquençage haut-débit (Illumina HiSeq 2000) pour caractériser les transcriptomes de souches isogéniques mâles et femelles *d'Ectocarpus* avec l'objectif d'identifier les gènes différentiellement exprimés chez les individus des deux sexes. Par ailleurs, le séquençage du génome d'un individu femelle a été réalisé et une ébauche d'assemblage a été réalisée par la Génoscope (Roche 454 + Illumina HiSeq 2000) dont l'un des objectifs est de disposer du locus sexuel spécifique à la femelle - SDR (Sex-determining Region) - qui a été identifié dans le génome mâle.

Pour caractériser les gènes spécifiques à la femelle, nous avons combiné l'utilisation des données génomiques et de données transcriptomiques. L'objectif est de permettre l'identification des scaffolds et des gènes spécifiques à la SDR femelle.

2 Matériel et Méthodes

Le séquençage du transcriptome a été réalisé par la société Fasteris (Suisse), sur des gamétophytes mâles et femelles matures, avec 2 réplicats biologiques et techniques pour chaque condition. Un nettoyage des séquences a été fait à l'aide du logiciel « FASTX-Toolkit », avec un filtrage des reads selon la qualité globale et un triming des reads selon la qualité de chaque pair de base. Le séquençage du génome femelle et son assemblage ont été réalisés par la Génoscope en combinant les technologies Roche 454 et Illumina. Le draft du génome a été assemblé en utilisant « Velvet » [2] pour la création des contigs et une étape de gap closing a été réalisée en utilisant « GapCloser » (« SOAPdenovo ») [3].

Une méthodologie d'assemblage de transcriptome *de novo* des réplicats femelle, utilisant Trinity [4], a été employée afin de d'obtenir un transcriptome exhaustif, comprenant les transcrits spécifiques au sexe femelle et les transcrits communs. L'identification des transcrits spécifiques à la femelle a été fait en réalisant un blast à la fois contre le génome mâle et contre le draft du génome femelle, en ne conservant que les « hits » ayants un match uniquement dans le génome femelle. Ces transcrits spécifiques à la femelle ont permis d'identifier des scaffolds potentiellement spécifiques de la SDR femelle. Une vérification par amplification PCR a été réalisée afin de confirmer les résultats.

La méthodologie d'assemblage de transcriptome avec un génome de référence a ensuite été réalisée à l'aide des logiciels « TopHat » et « Cufflinks » [5]. Le génome utilisé correspond à un « génome hybride », concaténation du génome de référence mâle et des scaffolds correspondant à la SDR femelle. Un premier assemblage a été réalisé en utilisant un fichier gtf, contenant les annotations des gènes caractérisés chez le mâle. Les résultats de cet assemblage ont été utilisés pour annoter les gènes spécifiques à la femelle et confirmer les résultats de l'annotation automatique de l'outil EuGene. Un second assemblage a été réalisé de la même manière que précédemment, mais en ajoutant les nouvelles annotations spécifiques à la femelle afin de calculer l'abondance des gènes spécifiques à la femelle avec une plus grande précision.

3 Résultats et Discussion

Le séquençage transcriptomique a généré en moyenne 26 millions de reads par réplicat. Le nettoyage des fichiers de reads a engendré une perte moyenne de 3 millions de reads pour chaque réplicat.

L'assemblage *de novo* des réplicats femelle a généré 60 000 transcrits dont 831 ont été identifiés comme spécifiques à la femelle. Ces transcrits spécifiques ont permis d'isoler dans l'assemblage du génome femelle, 97 scaffolds potentiellement spécifiques à la femelle. Après amplification par PCR, 34 scaffolds ont été confirmés comme appartenant à la SDR femelle.

Le premier assemblage avec référence réalisé avec les annotations du génome mâle a permis d'identifier 66 gènes potentiels dans la SDR femelle, avec une majorité de gènes mono-exonique (42). Une partie de ces résultats ont permis de confirmer la totalité des annotations automatiques de la SDR femelle et la découverte de nouveau gènes, avec au final 24 gènes annotés. Le second assemblage avec référence, réalisé en ajoutant les nouvelles annotations pour les gènes spécifiques à la femelle, a permis de recalculer l'abondance des gènes femelle avec une plus grande précision. Sur les 66 gènes potentiels identifiés lors du 1er assemblage, seul les 24 validés ont été conservés, la totalité des mono-exonique n'étant plus assemblés. La comparaison du nombre de reads mappés et leurs localisations dans la SDR femelle entre les deux assemblages a montré un recrutement préférentiel des reads dans les régions avec des annotations, expliquant la disparition des 42 gènes mono-exoniques.

Les approches d'assemblage *de novo* et avec référence se sont révélées complémentaires, permettant à la fois l'identification des scaffolds, la découverte et l'annotation des gènes d'un locus précis.

Acknowledgements

Financement du projet : Emergence UMPC 2010, ANR projet Blanc "Bi-cycle".

References

[1] J. M. Cock, L. Sterck, P. Rouzé, D. Scornet, A. E. Allen, G. Amoutzias, V. Anthouard, F. Artiguenave, J.-M. Aury, J. H. Badger, B. Beszteri, K. Billiau, E. Bonnet, J. H. Bothwell, C. Bowler, C. Boyen, C. Brownlee, C. J. Carrano, B. Charrier, G. Y. Cho, S. M. Coelho, J. Collén, E. Corre, C. Da Silva, L. Delage, N. Delaroque, S. M. Dittami, S. Ritter, S. Rousvoal, M. Samanta, G. Samson, D. C. Schroeder, B. Ségurens, M. Strittmatter, T. Tonon, J. W. Tregear, K. Valentin, P. von Dassow, T. Yamagishi, Y. Van de Peer, et P. Wincker, « The Ectocarpus genome and the independent evolution of multicellularity in brown algae », *Nature*, vol. 465, n^o 7298, p. 617‑621, juin 2010.

[2] D. R. Zerbino et E. Birney, « Velvet: algorithms for de novo short read assembly using de Bruijn graphs », *Genome Res.*, vol. 18, n^o 5, p. 821‑829, mai 2008.

[3] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam, et J. Wang, « SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler », *GigaScience*, vol. 1, n^o 1, p. 18, déc. 2012.

[4] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, et A. Regev, « Full-length transcriptome assembly from RNA-Seq data without a reference genome », *Nat Biotech*, vol. advance online publication, mai 2011.

[5] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, et L. Pachter, « Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks », *Nature Protocols*, vol. 7, n^o 3, p. 562‑578, mars 2012.

A Whole Genome Scan for *Plasmodium falciparum* Malaria : Genomewide Significant Linkage of Parasitemia and Mild Malaria with Chromosomes 6p21-p23 and 17p12

Audrey BRISEBARRE[1], Brice KUMULUNGUI[2] , Francis FUMOUX[2] and Pascal RIHET[1]

[1] TAGC, UMR1090 INSERM, Aix-Marseille Université, Parc scientifique de Luminy case 928, 163 avenue de Luminy, 13288, Marseille, Cedex 09, France

{brisebarre, rihet}@tagc.univ-mrs.fr

[2] UMR-MD3, Aix-Marseille Université, Campus de la Timone, 27 Bd Jean Moulin-CS30064, 13385, Marseille, Cedex 05, France

Keywords Malaria, genomewide linkage analyses, QTL.

1 Introduction

Human malaria caused by *Plasmodium falciparum* remains a major cause of mortality and morbidity worldwide. It has been shown that human genetic factors control the outcomes of *P. falciparum* malaria. Until now, two GWAS studies confirmed the HBB locus as a major locus in severe malaria. Linkage analyses pointed out some significant linkage on chromosomes 6p21-p23 and 10p15, and several suggestive linkage with mild malaria or parasitemia on chromosomes 2p25, 4q13-q21, 5p15-p13, 5q31-q33, 6p25.1, 6q15-q16, 12q21-q22, 13q13, 20p12 and 20q11 [1,2,3]. Furthermore linkage of mild malaria to chromosome 6p21-p23 and linkage of asymptomatic parasitemia to 5q31-q33 have been reported at least twice.

2 Results and Discussion

The present study reports results from a genome-wide scan and additional further testing of promising regions. We applied the maximum-likelihood binomial method extended to quantitative trait linkage analysis (MLB-QTL) and the quantitative trait disequilibrium test (QTDT) to search for genetic linkage and association with asymptomatic parasitemia and mild malaria. The initial study population consisted of 314 individuals belonging to 63 families living in Burkina Faso who were genotyped at 400 markers on autosomes, with an average distance of 10 cM [4]. Because chromosome 17p12 appeared to be very promising, 15 supplementary markers in this area were genotyped [4]. Additional individuals (n=247) that belonged to 55 families were also genotyped for the microsatellite markers [4].

Multipoint analyses revealed significant (LOD score >3.6) or suggestive (> 2.2) linkage. Indeed, quantitative trait linkage analysis confirmed significant linkage of mild malaria to chromosome 6p21.3 (LOD=3.73, p=0.000017), and provided suggestive linkage on chromosome 19p13.12 (LOD=2.50, p=0.000346) and of asymptomatic parasitemia to chromosomes 6p21.3 (LOD=2.35, p=0.000501) and 17p12 (LOD=2.77, p=0.00018) [figure 1]. Interestingly, the 1-LOD drop area for QTL location on chromosome 6p21.3 was very small, and included only 5 genes: TNF, LTA, LTB, LST1, and NCR3. The 1-LOD drop area for QTL location on chromosome 17p12 spanned 14cM, and contained 15 genes.

Because microsatellite markers on chromosome 17p12 were associated neither with mild malaria nor asymptomatic parasitemia, we further genotyped 4 additional SNPs within the 1-LOD drop area (rs9889434, rs13341772, rs9889936, and rs8077302). We did not detect any association with the SNPs. In contrast, haplotype analysis revealed a significant association with asymptomatic parasitemia (p=0.033).

Surprisingly, chromosome 5q31-q33 did not show linkage. We further tested association in presence of linkage between malaria phenotypes and microsatellite markers. After correcting for multiple tests with a FDR of 10%, there was an association of asymptomatic parasitemia with D5S642, the IL9 microsatellite, and D5S2017, which are genetic markers located within chromosome 5q31-q33.

Two chromosomal regions that we identified as carrying putative loci match to those identified in

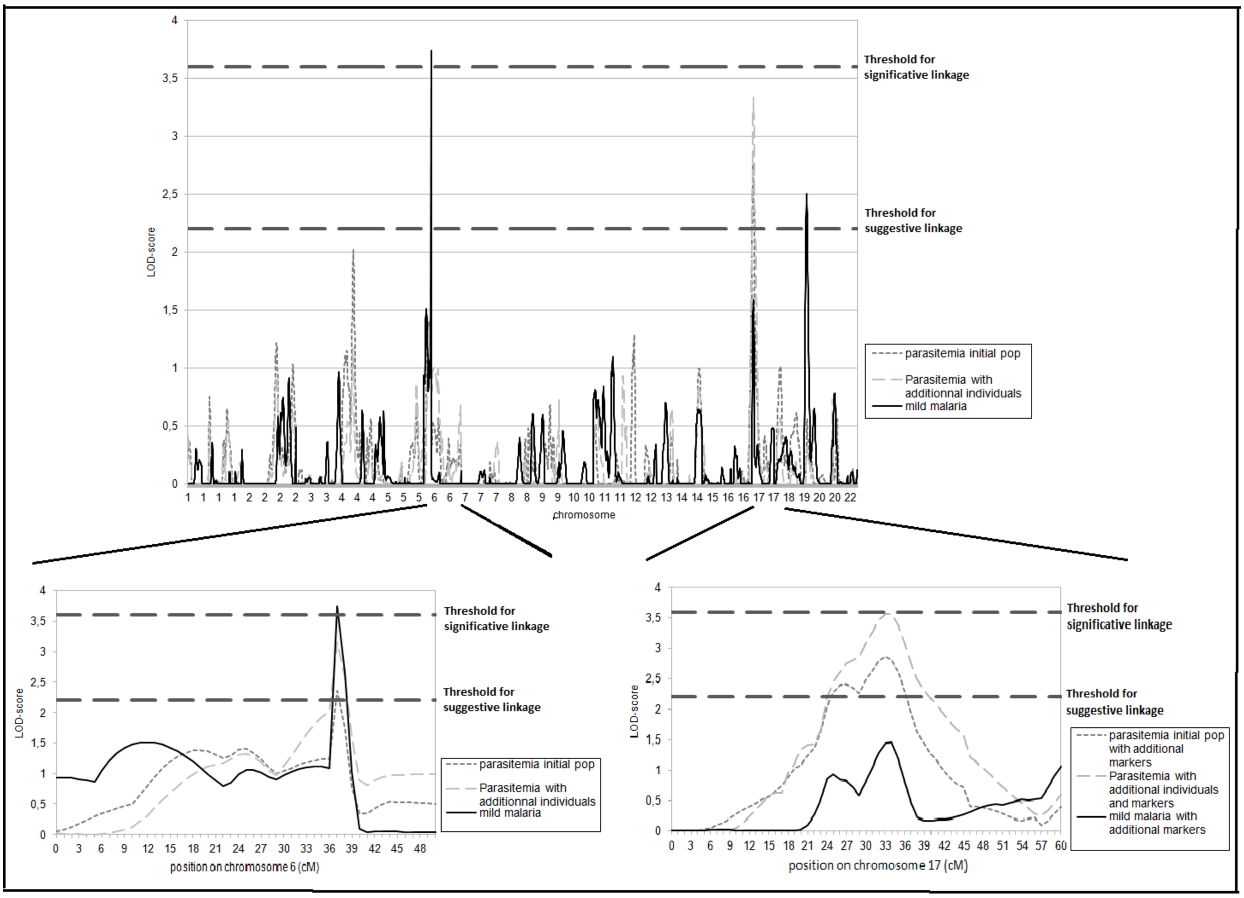


Figure 1. Genomewide scan of mild malaria and asymptomatic parasitemia. Multipoint LOD scores were plotted along the autosomes. Detailed views of the linkage peaks on chromosomes 6 and 17 are also plotted (2012).

other populations: chromosomes 5q31-q33 and 6p21-p23. In contrast, malaria phenotypes have not been previously linked to chromosomes 19p13 or 17p12. The linkage results give a complex picture of malaria resistance genetics. The analysis of phenotypes associated with mild malaria or parasitemia would contribute towards a better understanding of the genetics pathways involved.

Acknowledgements

This work was supported by grants from the French ministry (PhD fellowship), and Institut National de la Santé et de la Recherche Médicale (INSERM). We thank the Centre National de Genotypage for providing access to their genotyping facilities.

References

[1] J. Milet, G. Nuel, L. Watier, D. Courtin, Y. Slaoui, P. Senghor, F. Migot-Nabias, O. Gayeand A. Garcia, Genome wide linkage study, using a 250K SNP map, of Plasmodium falciparum infection and mild malaria attack in a senegalese population. *PloS ONE* 5(7):e11616, 2010.

[2] A. Sakuntabhai, R. Ndiaye; I. Casadémont, C. Peerapittayamonkol, C. Rogier, P. Tortevoye, A. Tall, R. Paul, C. Turbpaiboon, W. Phimpraphi, J-F. Trape, A. Spiegel, S. Heath, O. Mercereau-Puijalon, A. Dieye and C. Julier, Genetic determination and linkage mapping of Plasmodium falciparum malaria related traits in Senegal. *PloS ONE* 3(4):e2000, 2008.

[3] C. Timmann, J.A. Evans, I.R. König, A. Kleensang, F. Rüschendorf, J. Lenzen, J. Sievertsen, C. Becker, Y. Enuameh, K.O. Kwakye, E. Opoku, E.N.L. Browne, A. Ziegler, P. Nürnberg and R.D. Horstmann, Genome-wide linkage analysis of malaria infection intensity and mild malaria disease. *PloS Genet.*, 3(3):e48, 2007.

[4] P. Rihet, L. Abel, Y. Traore, T. Traore-Leroux, C. Aucan and F. Fumoux, Human malaria: segregation analysis of blood infection levels in a suburban area and a rural area in Burkina Faso. Genet. Epidemiol., 15:435-450, 1998.

IGO, Integration from Genomes to Organisms

Sandra DEROZIER[1], Hélène CHIAPELLO[2], Thomas LACROIX[1], Cyprien GUERIN[1], Valentin LOUX[1],
Anne GOELZER[1], Jean-François GIBRAT[1], Franck SAMSON[1], Annie GENDRAULT[1]

[1] UR 1077 Mathématique Informatique et Génome (MIG), INRA, Domaine de Vilvert, 78352, Jouy-en-Josas, France
{sandra.derozier, thomas.lacroix, cyprien.guerin, valentin.loux, anne.goelzer, jean-francois.gibrat, franck.samson, annie.gendrault}@jouy.inra.fr

[2] UR 0875 Mathématique et Informatique Appliquées Toulouse (MIAT), INRA, chemin de Borde-Rouge, 31326, Castanet-Tolosan, France
helene.chiapello@toulouse.inra.fr

Abstract *IGO portal is a web interface implemented with the Google Web Toolkit. It is a gateway to some of the resources developed at MIG laboratory. Currently five web interfaces and 389 complete prokaryotic genomes are integrated. IGO is accessible at http://migale.jouy.inra.fr/igo.*

Keywords genomic, integration, online resource.

IGO (Integration from Genomes to Organisms) is a web portal that interconnects different online tools developed at MIG laboratory in various domains : Mosaic [1] for the comparison of genomes, Prose [2] for the protein sequences, Bacillus subtilis Expression Data Browser [3] for the expression of B. Subtilis data, Pareo [4] for metabolic pathways and Insyght [5] for homologies/syntenies.

The integration of the different web tools relies on passing arguments as URL parameters. The benefit of this approach is that web applications from different technological backgrounds can be interconnected without major additional development. However, the web tools must support URL parameters such as the organism name, the strain, the gene name or the locus tag.

The IGO web interface has been developed using the Google Web Toolkit (GWT). Data are stored in a PostgreSQL relational database which currently contains 389 bacterial organisms.

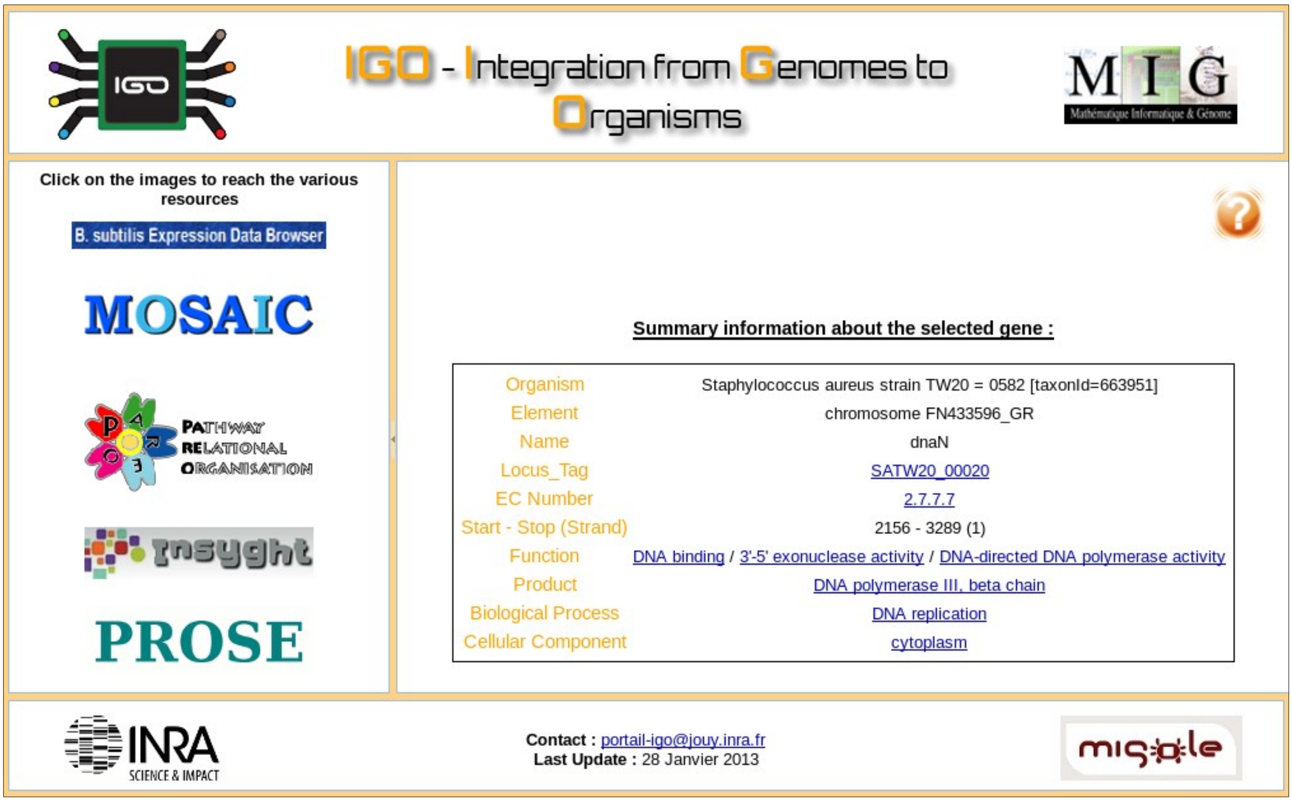
The database contains three categories of data :

- primary data such as genomic annotations that are extracted from EBI Genome Reviews files,
- secondary data corresponding to the cross comparisons, using BLAST, of all the genes of the stored bacterial genomes,
- tertiary data such as orthologs that are computed from the primary and secondary data.

The entry point in IGO is an organism, a strain and the chromosome or plasmid of interest. A table lists all the genes and various information such as the locus tag, genomic positions, EC number, product and Gene Ontology.

A search functionality is implemented to facilitate the retrieval of genes of interest. Different types of filters are available : genomic locations, presence/absence of homology, identifier (gene name/synonymous, locus tag, EMBL or Uniprot IDs), Gene Ontology terms (biological process, function, cellular component), EC number and product.

By selecting the gene name of interest, a summary of its information is available along with links pointing to the various resources developed at MIG laboratory (Fig. 1 shows an example).



IGO - Integration from Genomes to Organisms

Click on the images to reach the various resources

IGO
M I G
Mathématique Informatique & Génome
B. subtilis Expression Data Browser
MOSAIC
PAREO
PATHWAY RELATIONAL ORGANISATION
Insyght
PROSE
INRA
SCIENCE & IMPACT
migale

Summary information about the selected gene :

Organism	Staphylococcus aureus strain TW20 = 0582 [taxonId=663951]
Element	chromosome FN433596_GR
Name	dnaN
Locus_Tag	SATW20_00020
EC Number	2.7.7.7
Start - Stop (Strand)	2156 - 3289 (1)
Function	DNA binding / 3'-5' exonuclease activity / DNA-directed DNA polymerase activity
Product	DNA polymerase III, beta chain
Biological Process	DNA replication
Cellular Component	cytoplasm

Contact : portail-igo@jouy.inra.fr
Last Update : 28 Janvier 2013

Figure1. Summary information about the selected gene.

The IGO portal will be gradually enriched by integrating more MIG laboratory resources and all of available complete bacterial genomes.

Acknowledgements

The autors would like to thank Julien Jourde, Philippe Veber andLeslie Aïchaoui for their valuable comments.

References

[1] http://genome.jouy.inra.fr/mosaic/
[2] http://genome.jouy.inra.fr/prose/
[3] http://genome.jouy.inra.fr/cgi-bin/seb/index.py
[4] http://genome.jouy.inra.fr/pareo/
[5] http://genome.jouy.inra.fr/Insyght/

Genomics of adaptation during experimental evolution of legume symbionts

Camille Clerissi[1], Delphine Capela[1], Philippe Remigi[1], Rachel Torchet[2], Stéphane Cruveiller[2], Eduardo Rocha[3], and Catherine Masson-Boivin[1]

[1] INRA, LIPM, 24 Chemin de Borde Rouge, 31326, Castanet-Tolosan, France
{camille.clerissi, delphine.capela, philippe.remigi, catherine.masson}@toulouse.inra.fr
[2] Centre CEA/FAR – Institut de Génomique, 2 rue Gaston Crémieux, 91057, Evry, France
{scruveil, rtorchet}@genoscope.cns.fr
[3] Institut Pasteur, 25 rue du Docteur Roux, 75015, Paris, France
erocha@pasteur.fr

Abstract

Exchange of genetic material plays a major role in bacterium evolution. Among the best illustrations are nitrogen-fixing legume symbionts that evolved and spread in many unrelated phylogenetic branches through lateral transfer of essential symbiotic genes. The full phenotypic expression of the acquired traits may require readjustment of the new genetic background. However mechanisms allowing post-LGT adaptation are largely unknown.

To address this question, we took advantage of the experimental evolution of a pathogenic Ralstonia solanacearum chimera carrying the symbiotic plasmid of the rhizobium Cupriavidus taiwanensis into legume symbionts [1]. The chimeric Ralstonia was progressively adapted to nodule tissues by 9 parallel lineages of serial ex planta-in planta passages. Evolution was very fast, since the two first major symbiotic steps, nodulation and intracellular infection, have been acquired in less than 16 cycles [1,2].

Genome resequencing revealed an overabundance of mutations in our evolution experiment. A total of ca. 500 point-mutations were detected in the 9 final clones as compared to the original ancestor. We will present a first analysis of the genomic changes that came along with the adaptation process, i.e. mutation spectra, evolution of the number and nature of mutations and molecular convergences between lineages at the gene, operon and pathways levels.

Keywords

Comparative genomics, selection, convergence.

References

[1] Marchetti M. *et al.* (2010) Experimental Evolution of a Plant Pathogen into a Legume Symbiont. *PLoS Biology*, 8, doi:e100028010.1371/journal.pbio.1000280.

[2] Guan SH et al (2013) Experimental evolution of nodule intracellular infection in legume symbionts. *The ISME Journal*, doi:10.1038/ismej.2013.24

XML4NGS : A XML-based description of a Next-Generation sequencing project allowing the generation of a 'Makefile'-driven workflow.

Pierre LINDENBAUM[1], Raluca TEUSAN[2], Solena LE-SCOUARNEC[2], Audrey BIHOUÉE[2] and Richard REDON[2]

[1] CHU Nantes / UMR-1087, Institut du Thorax, 8 quai Moncousu, 44007, Nantes, France
pierre.lindenbaum@univ-nantes.fr @yokofakun
[2] UMR-1087, Institut du Thorax, 8 quai Moncousu, 44007, Nantes, France.

Abstract *XML4NGS is a schema describing a NGS experiment in XML. It provides a XSLT stylesheet transforming the XML into a Makefile-driven workflow allowing a parallel analysis (alignment, calling, annotation ...) on a cluster.*

Keywords NGS, XSLT, XML, workflow, pipeline, make, makefile, qmake, cluster, next-generation-sequencing.

XML4NGS : Une description en XML de projet de séquençage à haut débit permettant la génération de scripts analyse basés sur 'Make'.

Abstract *XML4NGS est un schéma décrivant une expérience de NGS en XML, le projet contient une feuille de style XSLT permettant de transformer ce document XML en fichier 'Makefile' permettant ainsi son analyse sur un cluster.*

Keywords XML, XSLT, NGS, pipeline, séquençage, cluster, make, qmake.

1 Introduction

Lors de projets de reséquençage d'exome ou de génome entiers, de nombreuses étapes sont indépendantes les unes des autres et peuvent être traitées en parallèle. Ainsi les alignements sur un génome de référence de deux fichiers FASTQ pour deux échantillons différents sont indépendants et peuvent être traités en parallèle. Un outil d'automatisation des tâches tel que le vénérable 'make'[3] est particulièrement adapté à ce genre d'analyse. En effet, l'option `--jobs` de make permet de spécifier le nombre de tâches parallélisables et peut donc aligner simultanément les FASTQs dans les limites de la machine. Par ailleurs, une version de make baptisée 'qmake'[2] est capable d'utiliser l'architecture d'un cluster et de distribuer les tâches sur les différents noeuds. Nous avons developpé un schéma en XML décrivant, entre autres, les échantillons et l'emplacement de leur fichiers FASTQs. Une feuille XSLT[1] permet de transformer ce XML en 'Makefile' et d'optimiser l'analyse du projet en parallèle.

2 Description

Le projet XML4Bio contient un utilitaire en Java qui va chercher des fichiers FASTQs dans une arborescence, extraire les noms des échantillons et préparer la description du projet en XML. Cette structure est par ailleurs spécifiée dans un document XML-schema/xsd. L'utilisateur a ensuite la possibilité d'éditer ce fichier XML afin de préciser les propriétés spécifiques du projet. Parmi les options disponibles se trouvent entres autres : le chemin d'accès vers le génome de référence, les étapes facultatives (recalibration, réalignement autour des indels), la manière de générer les fichiers VCF (un seul fichier par individu ou non), les annotations fonctionnelles à appliquer, etc... Une feuille de style XSLT est alors appliquée afin de transformer le fichier XML en fichier Makefile. Make (ou qmake) va ensuite exécuter l'enchaînement de commandes, uniquement si elles sont nécessaires, afin de produire des fichiers VCFs annotés à partir des fichiers FASTQs ainsi que des rapports de qualité. Make va également exploiter le fait que certaines tâches sont indépendantes afin de les paralléliser.

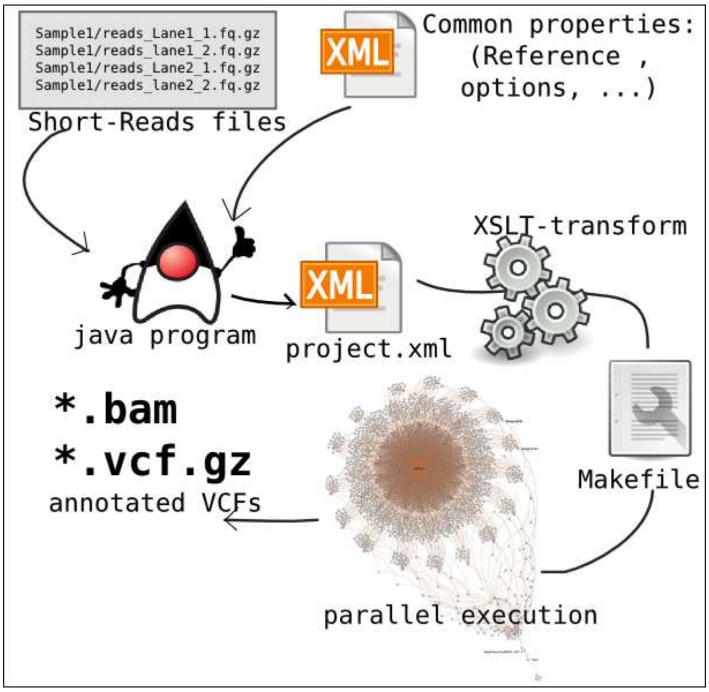

Sample1/reads_Lane1_1.fq.gz
Sample1/reads_lane1_2.fq.gz
Sample1/reads_Lane2_1.fq.gz
Sample1/reads_lane2_2.fq.gz
Short-Reads files
XML
Common properties: (Reference , options, ...)
java program
XML
project.xml
XSLT-transform
Makefile
*.bam
*.vcf.gz
annotated VCFs
parallel execution

Figure 1. Un programme java analyse la structure des répertoires contenant les FASTQs et génère un fichier XML décrivant le projet ainsi que ses propriétés. Ce fichier XML est ensuite transformé en fichier Makefile grâce à XSLT. La parallélisation de l'analyse est gérée par make/qmake et nous obtenons les BAMS ainsi que les fichiers VCFs annotés à la fin du processus.

3 Conclusion

Notre laboratoire utilise ce système XML+XSLT en production en conjonction avec un cluster SGE (Sun Grid Engine). Ce système nous donne pleinement satisfaction tout en réduisant le nombre d'opérations manuelles à effectuer.

4 Implémentation

Le projet est disponible sur github.com à l'URL suivante : `https://github.com/lindenb/xml4ngs`

Références

[1] http://en.wikipedia.org/wiki/XSLT

[2] http://gridscheduler.sourceforge.net/htmlman/htmlman1/qmake.html

[3] http://www.gnu.org/software/make/

Spatial statistical modelling of plant diversity from high-throughput environmental DNA sequence data

Angelika STUDENY[1], Florence FORBES[1], Celine MERCIER[2], Lucie ZINGER[2,3], Aurélie BONIN[2], Frédéric BOYER[2], Pierre TABERLET[2], Alain VIARI[1] and Eric COISSAC[2]

[1] INRIA Grenoble – Rhône-Alpes, 655 avenue de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France
[2] Laboratoire d'Ecologie Alpine, BP 53, 2233 rue de la piscine, 38041 Grenoble, Cedex 9, France
[3] Laboratoire Evolution et Diversité Biologique, Bâtiment 4R1 Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, Cedex 9, France
angelika.studeny@inria.fr

Abstract *This work aims at developing statistical models to investigate co-existence and spatial co-occurrence of species in an ecosystem. A special feature is the origin of the data from high-troughput environmental DNA sequencing of soil samples. We look at different modelling approaches, such as model-based clustering techniques, sparse matrix estimation and linear models of coregionalistation and their applicability to these data, with the aim of taking account of the spatial structure of the sampling design. Spatial cross-correlation in bivariate relationships is estimated. Different possible visualisations are discussed, in particular graph structures.*

Keywords Spatial correlation, Gaussian random fields, linear coregionalization, model-based clustering, environmental DNA, high dimensional data

Modélisation statistique spatiale de la diversité végétale basée sur des données haut-débit de la sequençage ADN

Résumé *Ce projet a comme objectif le développement de modèles statistiques pour étudier la co-existence et la co-occurrence spatiale des espèces dans un écosystème. L'originalité réside dans la particularité des données, génerées par du séquençage à haut-débit de l'ADN environnemental d'échantillons de sol. On étudie des approches différentes – des modèles de clustering spatial à base de mélanges, l'estimation des matrices de covariance sous régularisation ainsi que les modèles linéaires de corégionalisation, et leur aptitude face à ces données complexes avec l'objectif de prendre en compte le design spatial des echantillons. La corrélation spatiale dans des relations bivariées est estimée. Différentes possibilités de visualisation sont discutées, particulièrement celles sous forme des graphes.*

Mots-clés corrélation spatiale, champs aléatoires Gaussiens, modèles linéaires de corégionalisation, clustering, modèles de mélanges, ADN environnemental, données de haute dimension

1 Introduction

Molecular methods are currently revolutionizing traditional approaches in ecology of assessing species diversity, the composition of ecological communities and the co-occurrence of species. *Meta-barcoding* consists in extracting and identifying DNA-fragments from environmental samples with the help of genetic markers which are specific to different taxonomic groups [1]. In contrast to other methods, metabarcoding data contains indirect information on the presence of a target species. Thus they trace multiple taxa simultaneously across space. In analogy to DNA-barcoding, the genetic 'fingerprints' deposited in environmental samples, such as in soil, are amplified applying PCR (*polymerase chain reaction*) protocols and sequenced using high throughput techniques. They are subsequently identified at the lowest possible level in the taxonomic tree and distinguished as MOTUs (*molecular operating taxonomic units*) with the help of DNA-sequence reference databases, established on purpose or extracted from standard DNA libraries, such as *GenBank* (www.ncbi.nlm.nih.gov/genbank).

This paper presents potential modelling approaches of metabarcoding data with the aim of explicitly taking account of the spatial design of the survey to study species' distributions, to identify species co-occurrence patterns in space and subsequently to try inferring potential ecologically significant interactions between species.

2 Data and methods

Soil core samples have been taken on a grid of 19 × 19 points in a 100 ha square at the Nouragues CNRS Field Station in French Guiana. From these, extracellular DNA is isolated independently by PCR amplification and sequencing of the P6 loop of chloroplast *trn*L intron, which is discriminative for plant DNA [2]. Removing erroneous PCR output and sequences with insufficient data coverage, 342 MOTUs are distinguished as outline above. The data contain sequence read counts for each MOTU in each grid location. We seek methods that are applicable for biodiversity analysis of such high throughput environmental DNA data. A particular interest is in inferring *spatial* co-occurrence pattern between and within taxonomic groups. To this end, different approaches are studied and used in combination.

1. **Model-based clustering** can help reduce the dimension of the data by identifying those MOTUs which contain relevant signal as well as cluster similar observations in space. Dimension reduction is achieved by imposing sparsity constraints and locating observations in a lower-dimensional common subspace [3].
2. In a similar line, but without taking account of the spatial setting of the data, **sparse matrix estimation techniques** [4] allow us to reduce the dimension of the data and determine subgroups of MOTUs through independencies in the covariance matrix, visible in a graph structure.
3. **Linear models of coregionalisation** [5] are able to explicitly describe assumptions on the underlying spatial structure. Developed in a Bayesian framework, we use a fast, deterministic algorithm (based on integrated nested Laplace approximations `INLA`, [6]) to estimate a parametre of bivariate cross-correlation.

3 Preliminary results and their visualisation

Even with an efficient algorithm (`INLA`), computational costs for a relatively simple spatial model are high and have to be limited to bivariate relationships only. The power to detect co-occurrences is reduced when there is not a sufficient amount of data. The high dimensionality of the data complicates visualisation and ecological interpretation of the model output. Pooling MOTUs at higher taxonomic levels, we summarize possible interactions from positive and negative bivariate spatial correlations (FIG. 1(a)). Alternatively, subgroups are identified by imposing sparsity constraints on the correlation matrix. The resulting independencies between MOTUs can again be visualised by a graph structure with MOTUs as nodes (FIG. 1(b)). Cluster approaches on the other hand try to identify those samples which have a similar species composition (FIG. 1(c)).

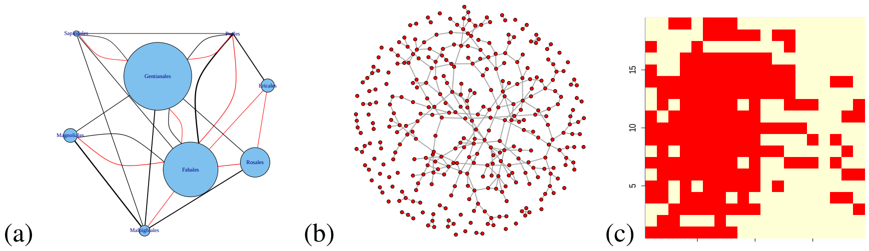


Figure 1. Visualising spatial co-occurrence of tropical plants by (a) graphs summarising positive (black) and negative (red) spatial interaction between members of different orders, (b) neighbourhood graphs representing the covariance matrix between MOTUs under sparsity constraints and (c) 2 spatial clusters of locations with similar species composition.

References

[1] L.S. Epp, S. Bossenkool, E.P. Bellemain, and et al. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. *Molecular Ecology*, 2012.

[2] P. Taberlet, E. Coissac, F. Pompanon, and et al. Power and limitations of the chloroplast *trnl*(uaa) intron for plant DNA barcoding. *Nucleic Acids Research*, 2007.

[3] C. Bouveyron and C. Brunet. Discriminative variable selection for clustering with the sparse Fisher-EM algorithm. *Computational Statistics*, 2013, In press.

[4] H. Liu and L. Wang. Tiger: A tuning-insensitive approach for optimally estimating Gaussian graphical models. Technical report, Massachusetts Institue of Technology, 2012.

[5] Alan E. Gelfand, Alexandra M. Schmidt, Sudipto Banerjee, and C.F. Sirmans. Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, 13:263–312, 2004.

[6] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B*, 71:319–392, 2009.

Repartition of β-CASP ribonucleases among Archaea

Petra S. LANGENDIJK-GENEVAUX[1,2] , Duy Khanh PHUNG[1,2], Dana RINALDI[1,2], Agamemnon J. CARPOUSIS[1,2], Béatrice CLOUET-D'ORVAL[1,2] and Yves QUENTIN[1,2]

[1] Laboratoire de Microbiologie et Génétique Moléculaires, UMR5100 CNRS, 118 Route de Narbonne, Bât. IBCG, 31062, Toulouse, Cedex 09, France

[2] Université de Toulouse, UPS, Laboratoire de Microbiologie et Génétique Moléculaires, F-31000, Toulouse, France

{yves.quentin, petra.genevaux, duy-khanh.phung, dana.rinaldi,agamemnon.carpousis, beatrice.clouet-dOrval}@ibcg.biotoul.fr

Abstract *Bacterial RNase J and eukaryal cleavage and polyadenylation specificity factor (CPSF-73) are members of the β-CASP family of ribonucleases involved in mRNA processing and degradation. Here we report an in-depth phylogenomic analysis that delineates aRNase J and archaeal CPSF (aCPSF) as distinct orthologous groups and establishes their repartition in 110 archaeal genomes. The genes of the aCPSF1 subgroup have been inherited vertically and are strictly conserved. These are characterized by an N-terminal extension with two K homology (KH) domains and a C-terminal motif involved in dimerization of the holoenzyme. Pab-aCPSF1 (*Pyrococcus abyssi *homolog) has an endoribonucleolytic activity that preferentially cleaves at single-stranded CA dinucleotides and a 5'-3' exoribonucleolytic activity that acts on 5' monophosphate substrates. These activities are the same as described for the eukaryotic cleavage and polyadenylation factor, CPSF-73, when engaged in the CPSF complex. Our results suggest that aCPSF1 performs an essential function and that an enzyme with similar activities was present in the last common ancestor of Archaea and Eukarya.*

Keywords Orthologs, nucleic acid enzymes, MCL, clustering, phylogeny, genomics.

1 Context and aim

RNA processing and degradation are critical to the survival of all cells and acknowledged as a means of regulating gene expression. In particular, the nature of RNA 5′ and 3′-ends is known to have major impact because they control the entry and directionality of endo- and exoribonucleases involved in these processes. In Archaea, exploration of RNA processing and degradation pathways is still in its early stages. The metallo β-lactamase protein superfamily is highly represented in Archaea [1,2]. Among them, the archaeal β-CASP family members were proposed to be RNA hydrolases as their sequences are more closely related to bacterial RNase J and to eukaryotic CPSF-73 than to β-CASP proteins involved in DNA repair and recombination [3,4]. The β-CASP proteins have the first four signature motifs of the metallo-β-lactamase superfamily followed by a distinct region (β-CASP domain, [3]) that is characterized by three short conserved motifs A (Asp or Glu), B (His) and C (His). These enzymes use a zinc-dependent mechanism in catalysis and act as 5′–3′ exonucleases and/or endonucleases [1-4]. The aim of this study is to establish the inventory, classification and phylogenetic analysis of the archaeal β-CASP proteins and to compare these with their counterparts involved in RNA metabolism in Bacteria and in Eukarya.

2 Methods and results

Genome entries of the complete archaeal and bacterial genomes were retrieved from EMBL (http://www.ebi.ac.uk/genomes/) and processed by a set of perl programs into a mySQL database. We used the β-CASP domain of each of 40 archaeal candidate sequences reported previously [3], as query in a Psi-Blast search against the protein sequences of 110 complete archaeal genomes (E-value < 1e-05, up to 20 iterations). This resulted in an initial collection of 375 proteins.

Pairs of one-to-one ortholog genes were computed with BlastP as follows: two genes a and b from genomes A and B, are considered to be orthologs if a is the best hit of b in genome A and reciprocally, and if a (or b) has a paralogous gene named c then the score of a versus b should be greater than the score of a (or b) versus c. For each protein of the initial collection, we retrieved orthologs in each archaeal genome to obtain orthologous proteins pairs. Among these, 13 proteins were not identified by the PsiBlast search. A graph was produced with the collected β-CASP sequences from 110 complete archaeal genomes, where vertices correspond to proteins and edges to orthologous relationships. The application of a partition algorithm (MCL [5]), revealed nine well-defined groups of proteins (orthologous groups, OGs). Members of the GloB and MtrA clusters did not have the conserved A, B and C motifs characteristic of β-CASP proteins; members of the αβCx cluster had non-canonical spacing between the A and B sequence motifs; members of the αβCy and αβCz clusters were not monophyletic, suggesting complex evolution that might include horizontal gene transfers with bacteria. Using the remaining four groups of orthologous β-CASP proteins, we constructed a tree that was rooted with eukaryotic CPSF-73 and bacterial RNase J. On this tree, bacterial and aRNase J were separated from the CPSF-like OGs by a long branch (100% bootstrap support), suggesting an early evolutionary separation. Proteins related to eukaryal CPSF-73 clustered into three OGs: aCPSF1 (112 members), aCPSF1b (11 members) and aCPSF2 (80 members). The members of aCPSF2 OG are distributed among Crenarchaeota, Euryarchaeota and Thaumarcheoata. The aCPSF1 OG corresponds to a highly conserved family with an N-terminal extension containing two KH RNA binding motifs specific to this group, and a C-terminal motif that is part of a protein dimer interface. This OG is notable because of its remarkable conservation in all Archaea with no exception to date. Moreover, the congruence between the archaeal and aCPSF1 phylogenetic trees shows that aCPSF1 has been inherited vertically, suggesting an ancient origin predating the emergence of Archaea. The small aCPSF1b OG branching close to the aCPSF1 OG is restricted to the Methanococcales. The aCPSF1b proteins, which appear to have an undecipherable ancient origin, lack the N-terminal extension that is characteristic of the aCPSF1 family.

Pab-aCPSF1 (*Pyrococcus abyssi* homolog) has an endoribonucleolytic activity that preferentially cleaves at single-stranded CA dinucleotides and a 5'-3' exoribonucleolytic activity that acts on 5' monophosphate substrates. These activities are the same as described for the eukaryotic cleavage and polyadenylation factor, CPSF-73, when engaged in the CPSF complex. The N-terminal KH domains are important for endoribonucleolytic cleavage at certain specific sites and the formation of stable high molecular weight ribonucleoprotein complexes. Dimerization of Pab-aCPSF is important for exoribonucleolytic activity and RNA binding.

In conclusion, our results suggest that aCPSF1 performs an essential function in RNA metabolism and that an enzyme with similar activities was present in the last common ancestor of Archaea and Eukarya.

References

[1] Z. Dominski, Nucleases of the metallo-beta-lactamase family and their role in DNA and RNA metabolism. *Crit. Rev. Biochem. Mol. Biol.*, 42:67-93, 2007.

[2] B. Mir-Montazeri, M. Ammelburg, D. Forouzan, A.N. Lupas and M.D. Hartmann, Crystal structure of a dimeric archaeal cleavage and polyadenylation specificity factor. *J. Struct. Biol.*, 173:191-195, 2011.

[3] I. Callebaut, D. Moshous, J.P. Mornon and J.P. de Villartay, Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family. *Nucleic Acids Res.*, 30:3592-3601, 2002.

[4] B. Clouet-d'Orval, D. Rinaldi, Y. Quentin and A.J. Carpousis, Euryarchaeal beta-CASP proteins with homology to bacterial RNase J Have 5′- to 3′-exoribonuclease activity. *J. Biol. Chem.*, 285:17574-17583, 2010.

[5] S. van Dongen and C. Abreu-Goodger, Using MCL to extract clusters from networks. *Methods Mol. Biol.*, 804:281-295, 2012.

Computational analysis and modeling of the genetic regulatory network controlling rhombomere development.

Daniela GARCIA[1], Morgane THOMAS-CHOLLIER[1], Pascale GILARDI-HEBENSTREIT[2], Patrick CHARNAY[2] and Denis THIEFFRY[1]

[1] Computational systems biology, Ecole Normale Supérieure, IBENS, CNRS 8197, INSERM 1024,
46 rue d'Ulm, 75230 Paris cedex 05, France
dgarcia@biologie.ens.fr , mthomas@biologie.ens.fr , thieffry@ens.fr

[2] Development of the nervous system, Ecole Normale Supérieure, IBENS, CNRS 8197, INSERM 1024,
46 rue d'Ulm, 75230 Paris cedex 05, France
gilardi@biologie.ens.fr , charnay@biologie.ens.fr

Keywords Krox20, logical modeling, pattern-matching, cis-regulatory elements

Krox20 (encoded by the gene Egr2) is a key transcription factor implicated in hindbrain development, in particular during the formation of segments along the neural tube called rhombomeres (r). Krox20 expression in rhombomeres r3 and r5 is under the control of three main enhancer elements named A, B and C, located more than 200kb upstream in mouse and human, and 80kb upstream in chick and zebrafish (see [1], complemented by personal data]. Elements B and C are conserved across these species, whereas the auto-regulatory element A is only conserved in mouse, human and chick. Yet, we have recently localized a counterpart of element A in zebrafish, using available H3K4me1 ChIP-seq data (these marks are associated with cis regulatory elements), in silico identification of Krox20 binding sites, along with in vivo reporter gene experiments. It nevertheless remains unclear how this element is functionally acting as element A despite its lack of sequence conservation.

Our first goal was thus to define whether the sequence of zebrafish element A contains an auto-regulation signature despite its lack of global conservation across different vertebrate species. To address this problem, we first refined the Krox20 consensus matrix using known Krox20-regulated enhancers. Then, we analyzed the organization of element A Krox20 binding sites in clusters and compared this organization between zebrafish and tetrapods elements A using dotplots. For the pattern-matching approach, we used higher-order background models with the program *matrix-scan* from the RSAT software suite [2-4] to reduce the number of false positive predictions. We further predicted putative cofactors by using motif discovery and motif enrichment approaches with collections of binding motifs from the TRANSFAC and Jaspar databases.

Next, we compared the cis-regulatory organization of element A to the other Krox20 auto-regulatory elements (acting in neural crest, bone forming cells) and direct Krox20-regulated enhancers acting in rhombomeres 3 and 5 (*Hoxb2, Hoxa2, Hoxb3, EphA4*). We wonder whether the tissue specificity of these enhancers is achieved by different Krox20 cofactors. Preliminary results show as an example that with this approach, we were able to recover putative binding-sites for Mafb, which has been previously reported as a Krox20 cofactor [5]. Other potential transcription factors identified include Jun and Fox family members..

Finally, we built a dynamical logical model of the genetic regulatory network involving *Krox20* and *Hoxb1* genes in hindbrain patterning. These two genes auto-regulate and cross regulate each other. More precisely, Hoxb1 can both directly activate (in r3) [6] and indirectly repress (in r4) *Krox20* by Nlz (personal data). We are using the software GinSIM [7] to model these complex relationships; which has already provided insights regarding the mechanisms and dynamic interactions between the involved components (genes and enhancers). We have run various simulations starting with different initial states (e.g. corresponding to r3, r4 and r5 situations at early stages) and considering different genetic backgrounds (e.g. with knock-out or ectopic expression of components of the system) and compared the final fates to current data. Interestingly, this analysis suggests a hierarchy between the repressing effects of Nlz onto Krox20 via elements A and C, which remains to be tested experimentally.

Altogether, our detailed analysis of Krox20 cis-regulatory regions and the dynamical modeling of the corresponding regulatory network should allow us to better understand the mechanisms that control Krox20 expression during hindbrain development in vertebrates.

References

[1] Chomette, D., Frain, M., Cereghini, S., Charnay, P. and Ghislain, J (2006). Krox20 hindbrain cis-regulatory landscape: interplay between multiple long-range initiation and autoregulatory elements. *Development*,**133**, 1253-62.

[2] van Helden J (2003). Regulatory sequence analysis tools. *Nucleic Acids Res* **31**, 3593-6.

[3] Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E, Brohe, S, van Helden J (2008). RSAT: regulatory sequence analysis tools. *Nucleic Acids Res* **36**, W119-27, 2008.

[4] Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J (2011). RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res* **39**, W86-91.

[5] Manzanares,M, Nardelli J, Gilardi-Hebenstreit P, Marshall H, Martinez-Pastor MT, Krumlauf R, Charnay P (2002). Krox20 and kreisler co-operate in the transcriptional control of segmental expression of hoxb3 in the developing hindbrain. *EMBO J* **21**, 365-76.

[6] Wassef MA, Chomette D, Pouilhe M, Stedman A, Havis E, Desmarquet-Trin Dinh C, Schneider-Maunoury S, Gilardi-Hebenstreit P, Charnay P, Ghislain J (2008). Rostral hindbrain patterning involves the direct activation of a Krox20 transcriptional enhancer by Hox/Pbx and Meis factors. *Development* **135**, 3369-78.Chaouiya C, Naldi A, Thieffry D (2012). Logical modelling of gene regulatory networks with GINsim. *Meth Mol Biol* **804**, 463-79.

Logical modelling of immune cell specification and reprogramming

Samuel Collombet[1], Thomas Graf[2], and Denis Thieffry[1]

[1] IBENS (CNRS UMR 8197 / INSERM U1024), Paris, France
{samuel.collombet, thieffry}@ens.fr

[3] Center for Genomic Regulation, Barcelona, Spain
{thomas.graf}@crg.eu

Keywords: hematopoiesis, trans-differentiation, regulatory network, logical modelling.

Immune cells arise from a common set of multipotent progenitors, which can be driven by external stimulations into specific lineages like myeloid and lymphoid lineages. Haematopoietic cell specification involves different signalling pathways and various transcription factors. Gain- or loss-of-functions experiments have shown that some factors are necessary and/or sufficient to induced differentiation of progenitors, including EBF and Pax5 for B cells, or PU1 and Egr1 for macrophages [1]. Some of these factors can even force cells committed to a lineage to trans-differentiate into another one [2]. They form an intertwined regulatory network and the effect of a given factor often depends on the presence of co-factors [3].

The integration of existing data into a formal model allows to test their consistency and to predict the effects of perturbations *in silico*. The logical modelling framework can be used to model large regulatory networks and to simulate their dynamic in a qualitative way, at a low computational cost. Novel methods and tools can be used to link dynamical properties and topological characteristic of the network (analysis of functional circuit, of key transitions, of paths between different states, etc. [4, 5]). This formalism as proved to be useful for the study of related developmental problems [6].

Using data from molecular genetic, functional genomic (microarray, ChIP-seq [1, 2]) and DNA sequence analysis, we have build a model focusing on the regulatory networks underlying the specification of B cells and macrophages from common progenitors.

A first round of simulations led us to identify poorly documented regulatory interactions, in particular regarding the effect of B cell factors (EBF and Pax5) on the macrophages factors Cebpa and PU1. We are currently investigating these regulations by ectopic expression and knock-down experiments, using retro-/lenti-virus and shRNA.

Our modelling work has already contributed to identified caveats in the current view of haematopoiesis regulatory network and should enable us to refine our model. In parallel, we are currently expanding our model to encompass T cell specification, along with alternative myeloid lineages.

References

[1] C. Spooner, J. Cheng, E. Pujadas P. Laslo and H. Singh, A recurrent network involving the transcription factors PU.1 and Gfi1 orchestrates innate and adaptive immune cell fates. *Immunity* 31: 576-586, 2009.

[2] A. Di Tullio, T. Phong Vu Manh, A. Schubert, R. Mansson and T. Graf CCAAT/enhancer binding protein alpha (CEBPa)-induced transdifferentiation of pre-B cells into macrophages involves no overt retrodifferentiation. *Proceedings of the National Academy of Sciences* 108: 1–6, 2011

[3] M. Zarnegar, E.V. Rothenberg, Ikaros represses and activates PU.1 cell-type-specifically through the multifunctional Sfpi1 URE and a myeloid specific enhancer. *Oncogene* 31: 4647–4654, 2012.

[4] C. Chaouiya, A. Naldi, D. Thieffry, Logical modelling of gene regulatory networks with GINsim. *Methods in Molecular Biology* 804: 463-479.

[5] D. Bérenguier, C. Chaouiya, P.T. Monteiro, A. Naldi, E. Remy, D. Thieffry, L. Tichit, Dynamical modeling and analysis of large cellular regulatory networks, *Chaos*, accepted.

[6] A. Naldi, J. Carneiro, C. Chaouiya, D. Thieffry, Diversity and plasticity of Th cell types predicted from regulatory network modelling. *PLoS Computational Biology* 6: e1000912, 2010.

FEELnc: Fast and Effective Extraction of Long non-coding RNAs

Thomas DERRIEN[1], Catherine ANDRE[1] and Christophe HITTE[1]
CNRS UMR6290 , Institut de Génétique et Développement de Rennes , Université Rennes1, 35000 Rennes, France
{tderrien, candre, hitte}@univ-rennes1.fr

Abstract
Whole transcriptome sequencing technology (RNASeq) has become a standard for transcriptome analysis and allows to catalogue the RNA populations of a given cell line, a tissue at different time points and conditions. Amongst all RNAs, the class of the long non-coding RNAs (lncRNAs) is now emerging as a major component of the non-coding transcriptome. Indeed, several studies have shown their crucial roles in the cell machinery such as gene regulation, imprinting, X-inactivation. Annotating and classifying lncRNAs using RNASeq experiments remains a challenging task. Therefore, we present FEELnc a workflow for extracting and annotating lncRNAs based on a set of assembled transcripts as input. First, the program filters out transcripts that more likely correspond to assembly artifacts, then it checks the non protein-coding potential of each transcript and finally classifies lncRNAs based on their genomic localization in comparison to a reference set of protein coding genes. We applied FEELnc on RNASeq datasets from the dog species and identified more than 20,000 lncRNAs. FEELnc is a novel workflow that outputs a set of high-quality lncRNAs and can easily be integrated in RNASeq post-processing analysis.

Keywords lncRNA, NGS, RNASeq, non-coding, dog.

1 Introduction

The transcriptome of a cell is represented by a myriad of different RNAs molecules with and without coding capabilities [1]. Recent advances in transcriptome sequencing (RNASeq) have led to the identification of a new class of non-coding RNAs, the long non-coding RNAs (lncRNAs) [2-4]. LncRNAs are transcripts longer than 200 nucleotides with similar properties to protein-coding RNAs (mRNAs) i.e splicing, polII transcription, polyadenylation but without a functional open reading frame (ORF). Past studies on known lncRNAs have shown that they are involved in important molecular functions such as imprinting [5], X-inactivation [6] or more recently in regulation of carcinogenesis [7] or pluripotency [8].

Following RNA sequencing, one strategy consists in mapping the sequences or reads back onto the reference genome and reconstruct transcript models using dedicated tools [9, 10]. Although RNASeq has a very low background signal compared to hybridization-based approaches [11], it remains crucial to discriminate *bona fide* novel transcripts (such as putative lncRNAs) from transcript artifacts and/or background transcription. To this end, we present an integrative tool called FEELnc which uses a set of newly assembled transcripts as input and allows users to rapidly extract and annotate long non-coding RNAs. The program applies a series of customizable filters on the assembled transcripts (transcript length, mRNAs overlap, number of exons…), computes the coding potential using three complementary tools and classifies new lncRNAs according to their genomic localization compared to a set of reference protein-coding genes (intergenic, intragenic, exonic, intronic…). We used FEELnc on a set of 58 RNASeq to identify lncRNAs of the domestic dog genome.

2 FEELnc workflow

2.1 Customizable filtering

Given a reference genome annotation and a set of newly assembled transcripts in BED or GTF formats, FEELnc first discards any transcripts overlapping a reference protein-coding exon on the same strand. A default size filtering step (fixed at 200bp by definition) removes transcript shorter than this cutoff. Others tunable filters include the removal of (i) mono-exonic transcripts as they might correspond to mapping artifacts due to repetitive sequences for example, (ii) bi-exonic transcripts harboring one very short exon (<10bp by default) which could reflect a dubious assembled transcript and (iii) transcripts lying in coding loci if the user interest is to focus on intergenic lncRNAs (lincRNAs).

2.2 Non protein-coding potential and genomic classification

After the initial filtering phase, FEELnc employs three complementary programs, CPC [12], TxCdsPredict [13] and CPAT [14], to compute the coding potential of the remaining transcripts. A transcript is defined as lncRNA when flagged as non protein-coding by the three tools. A graphical representation is provided by a Venn diagram of the three sets so that the user can, for instance, relax method-specific parameters in order to visualize the distribution of lncRNAs in the intersection of the programs or in each program taken individually. Finally, given a reference annotation, FEELnc provides a lncRNA classification based on their distance and orientation (sense or antisense) with respect to the protein-coding set. It classifies (i) intergenic lncRNA (lincRNAs) and informs on the distance and orientation with the closest mRNAs (divergent, convergent or same strand) and (ii) FEELnc classifies intragenic lncRNAs based on their overlap with mRNAs exons (exonic antisense category) or introns (intronic sense and antisense). Theses lncRNA-mRNA classes could be further used to extract lncRNA candidates as regulator of its mRNA partner.

3 Application to the dog transcriptome

We used 58 canine RNASeq dataset from 15 tissues mapped on the dog reference genome with tophat2 suite [3]. FEELnc annotated ~20,000 multi-exonic lncRNAs, longer than 200 bp (mean size=1,172 bp) without protein-coding capabilities. Most of these canine lncRNAs are localized in intergenic regions (~90%) and represent good candidates for *cis*-regulators of neighboring mRNAs. Thus, FEELnc facilitates the annotation of high-confidence long-non coding RNAs and paves the way for further bioinformatic and/or experimental functional analysis.

Acknowledgements

We thank the Genouest plateform, the Genscale (IRISA) and “Canine genetics” (IGDR) teams. We also thank Anne-Sophie Guillory, Edouard Cadieu and Benoit Hédan for canine RNA extractions and experimental design.

References

[1] S. Djebali, C.A. Davis, A. Merkel, et al, *Nature*, 488:101–108, 2012.

[2] M. Guttman, I. Amit, M. Garber, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458:223–227, 2009.

[3] M.N. Cabili, C. Trapnell, L. Goff, et al., Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* 2011.

[4] T. Derrien, R. Johnson, G. Bussotti, et al., The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res*, 22:1775–1789, 2012.

[5] T. Nagano, P. Fraser, No-Nonsense Functions for Long Noncoding RNAs. *Cell*, 145:178–181, 2011.

[6] N. Brockdorff, A. Ashworth, G.F. Kay,et al., The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, 71:515–526, 1992.

[7] O. Wapinski, H.Y. Chang, Long noncoding RNAs and human disease. *Trends in Cell Biology*, 21:354–361, 2011.

[8] M. Guttman, J. Donaghey, B.W. et al., lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* :1–11, 2011.

[9] C. Trapnell, B.A. Williams, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28:511–515, 2010.

[10] S.B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, et al., Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464:773–777, 2010.

[11] Z. Wang, M. Gerstein, RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10:57–63, 2009.

[12] L. Kong, Y. Zhang, Z.Q. Ye, X.Q. Liu, S.Q. Zhao, L. Wei, G. Gao, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*, 35:W345–W349, 2007.

[13] B. Rhead, D. Karolchik, et al., The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*, 38:D613–9, 2010.

[14] L. Wang, H.J. Park, S. Dasari, S. Wang, J.P. Kocher, W. Li, CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* 2013.

Archive: structure and preserve data and metadata.

Ludovic LEGRAND, Ludovic COTTRET, Emmanuel COURCELLE, Erika SALLET, Sébastien CARRERE and Jérôme GOUZY.

Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR INRA-CNRS 441/2594, F-31320 Castanet Tolosan, France

{Ludovic.Legrand, Ludovic.Cottret, Emmanuel.Courcelle, Erika.Sallet, Sebastien.Carrere, Jerome.Gouzy}@toulouse.inra.fr

Keywords data repository, metadata, data preservation

Archive : structurer et préserver données et métadonnées.

Mots-clés référentiel de données, métadonnées, préservation des données

1 Introduction

Aujourd'hui, la pérennité des données brutes issues des sciences de la vie est assurée de façon très inégale puisque placée sous la responsabilité des laboratoires/instituts qui les produisent. Elle est pourtant indispensable pour assurer sur le long terme la traçabilité et l'exploitation de ces données aussi bien avec les méthodes et outils actuels que futurs. La diversité des données et des formats, l'évolution rapide des technologies à haut débit ainsi que la quantité croissante de données nécessite la mise en place d'un système à la fois robuste et générique qui permettra d'exploiter les données plusieurs années après leur production.

Pour répondre à ce besoin critique, nous proposons Archive, un système d'information dont les objectifs sont multiples :

- conservation pérenne (10 ans et plus) des données expérimentales brutes
- flexibilité pour gérer tout type de données expérimentales
- partage sécurisé des données
- facilité d'accès aux données par interface web et par accès programmatique

2 Métadonnées

Afin de pouvoir réutiliser les données expérimentales, leur description est indispensable pour garder une trace des caractéristiques des échantillons, des conditions de production et des technologies utilisées. Dans Archive, les métadonnées sont stockées dans des fichiers XML spécifiques à chaque type de données. Dans le cas de données de séquences nous nous sommes basés sur les métadonnées de la banque Gene Expression Omnibus (GEO) [1] et de SRA [2].

Le format XML est un format texte ouvert, et donc pérenne sur une durée longue. De plus, sa structure nous permet d'assurer la cohérence des informations à partir de schémas XSD tout en offrant la possibilité d'exporter les métadonnées vers d'autres formats grâce à la transformation XSLT. En plus des validations effectuées au niveau XML, un contrôle d'intégrité des données est effectué par un programme spécifique du type de données considéré (ex : détection de corruption de fichiers ou d'incohérences entre métadonnées et fichiers de données). Seules les données et métadonnées satisfaisant aux deux niveaux de contrôles sont effectivement copiées dans l'Archive.

3 Interface utilisateurs

Une interface web permet aux utilisateurs d'ajouter et de gérer facilement les données et les métadonnées. L’intégration des données dans Archive se fait en deux étapes : dans un premier temps il est nécessaire de transférer les données soit par HTTPS soit par SFTP puis de saisir les métadonnées associées aux fichiers à l'aide d'un formulaire généré à partir du XML décrivant ce type de données/métadonnées. Ainsi la cardinalité attendue, le caractère obligatoire ou optionnel, les énumérations, etc. sont décrits dans le XML et sont utilisés tant pour valider le fichier que pour générer les interfaces de saisies.

L'interface de consultation des données permet de visualiser les métadonnées, de télécharger les données et d'effectuer des recherches grâce au moteur d’indexation/recherche Lucene [3].

4 Authentification, droits d’accès et publication

La sécurisation de l'accès aux données est tout aussi importante que leur pérennité. L'application utilise la Fédération Éducation-Recherche mise en place au niveau national par Renater [4] pour authentifier les utilisateurs, et utilise le logiciel Shibboleth [5]. L'appartenance de l'utilisateur à un ou plusieurs groupes, ainsi que les rôles que l'administrateur d'Archive lui a donné sont utilisés pour contrôler son accès aux données privées. Tous les échanges se font par des protocoles sécurisés (HTTPS et SFTP) afin de garantir la sécurité des données privées. Archive fournit un export des métadonnées au format DublinCore et permet si nécessaire de publier les données avec le système Digital Object Identifier (DOI).

5 Architecture physique

En plus de la description des données, le stockage des données sur le long terme nécessite la mise en place de mécanismes permettant d'assurer la conservation des données et une reprise d'activité rapide en cas d'incident. L'application supporte le stockage sur un système de fichiers classique, mais aussi sur une grille de données iRODS [6]. Le premier choix nécessite la mise en place de sauvegardes alors que le second permet de répliquer les données à distance entre plusieurs serveurs (éventuellement installés sur plusieurs sites) de manière transparente. Le système iRODS, utilisé essentiellement par la communauté des physiciens, permet en effet de gérer de très grosses volumétries avec des systèmes de stockage hétérogènes, tout en assurant leur réplication.

6 Accès programmatique

Une interface programmatique sécurisée et basée sur REST ainsi qu'une bibliothèque écrite en Perl rend l'application interopérable avec les outils d'analyse, qui pourront ainsi utiliser les données stockées dans Archive.

La bibliothèque Perl offre la possibilité d’interroger plusieurs instances d'Archive si celles-ci sont enregistrées dans l'annuaire.

Enfin un client en ligne de commande utilisant cette bibliothèque permet de récupérer l'ensemble des données accessibles à partir d'une requête, soit par téléchargement soit par des liens symboliques en local.

Depuis février 2013 les données de séquences (fastq et sff) produites au Laboratoire des Interactions Plantes-Microorganismes (LIPM) sont progressivement intégrées au système d'information Archive. Les prochaines étapes concernent l'ajout du support de données phénotypiques et métaboliques, et en collaboration avec le réseau INRA BioInformatique, Biodiversité, Représentation et Intégration des Connaissances (BBRIC), la mise en place d'une réplication des données multi-sites, ainsi que l’interconnexion entre Archive et les moteurs de workflows.

Le code source de l’application est disponible à l’adresse suivante : http://lipm-svn.toulouse.inra.fr/svn/inra_archive

Remerciements

Nous remercions Valentin Loux et les membres du réseau INRA BBRIC pour leurs évaluations et leurs suggestions.

References

[1] Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo

[2] Sequence Read Archive. http://www.ncbi.nlm.nih.gov/sra

[3] Lucene project. http://lucene.apache.org/core/

[4] Fédération d'identité Education-Recherche. https://services.renater.fr/federation/index

[5] Shibboleth project. http://shibboleth.net/

[6] iRODS project. https://www.irods.org/index.php

In silico design and implementation of synthetic metabolic pathways in microbial cells factories

Gilles VIEIRA[1,2,3], Stéphanie HEUX[1,2,3] and Jean-Charles PORTAIS[1,2,3]

[1] Université de Toulouse; INSA, UPS, INP; LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France,
[2] INRA, UMR792, Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France,
[3] CNRS, UMR5504, F-31400 Toulouse, France
vieira@insa-toulouse.fr

Abstract *The use of renewable carbon sources for cost-effective, microbial production of compounds of interest has become a major issue of white biotechnology in the context of environmental concerns. To this aim, the rational optimization and modification of the metabolism is a central point. Micro-organisms, like E. coli and S. cerevisiae are widely used as hosts for implementing new metabolic functions. To construct these cell factories one need*
(i) to be able to design new pathways, that respect biological constraints;
(ii) to optimize the natural host metabolic network in order to derive the fluxes for a specific aim without impacting growth and robustness;
(iii) to understand the effects of such modifications, either on local metabolism (metabolic pathways) or on the entire metabolism of the host.
To achieve these objectives, we have developed a general framework for the metabolic design of synthetic pathways, fluxes optimization and pathway-interaction investigations. Using both experimental data from E. coli and S. cerevisiae, and in silico modeling approaches, here we present a generic workflow enabling the use of a novel carbon source to the overproduction of a specific metabolite by optimizing its precursors availability. We also illustrate how pathway-interaction investigations can help in understanding the global metabolism behavior after the introduction of natural and/or synthetic metabolic pathways.

Keywords Metabolic model, pathway design, metabolic optimization, pathway interactions

RNA-Seq data analysis with a focus on the human Major Histocompatibility Complex (MHC) applied to type 1 diabetes (T1D)

Azadeh SAFFARIAN[1,2], Julian KNIGHT[3], Patrick CONCANNON[4], Cécile JULIER[1,2] and Claire VANDIEDONCK[1,2]

[1] INSERM UMR-S 958, F-75010 Paris; [2]Univ Paris Diderot, Sorbonne Paris Cité, F-75013 Paris; [3]Wellcome Trust Centre for Human Genetics, Oxford University, UK; [4] University of Florida, Genetics Institute, Gainesville, USA

azadeh.saffarian@inserm.fr, claire.vandiedonck@inserm.fr

Keywords Strand-specific RNA-seq, Major Histocompatibility Complex (MHC), mapping, Human Leukocyte Antigen (HLA), allele-specific expression

1 Introduction

The human Major Histocompatibility Complex (MHC) region is known to play the largest role in Type 1 Diabetes (T1D) genetic predisposition. However only half of its genetic contribution is explained by Human Leukocyte Antigen (HLA) class II loci and other HLA and non-HLA MHC genes must been involved. Our overall objective aims at identifying new T1D susceptibility genes by characterizing genes differentially expressed between T1D patients and their familial controls, with a major focus on the (MHC) region. This region being the most gene dense region of the human genome with numerous overlapping genes on both strands, the chosen approach is a directional RNA-Seq to produce strand-specific information following the dUTP protocol [1]. It is performed with an Illumina Hi-Seq 2000 technology with >15x10^6 paired-end reads of 2x100 bp. A pilot study to evaluate the directionality and allele specificity is in progress to validate the protocol, assess the strand specificity and confirm allele-specific expression. Our specific aims are: to set up an analysis pipeline from raw data to the identification of differentially expressed genes; and to develop a further algorithm for mapping paired-end reads to the highly polymorphic classical HLA genes.

2 Pilot Study

We used the same three MHC homozygous samples, PGF, COX and QBL, as in [2] were used. To assess allele specific expression, we mixed PGF and COX RNAs at different ratios. After removal of rRNAs, dUTP strand-specific libraries were prepared. All samples were pooled and sequenced twice on two different lanes of a same flow-cell with a final depth of 17 to 27.7 millions 100-bases paired-end reads per sample using the Illumina Hi-Seq 2000 technology. In addition to standard quality control process, we verified both ends of a pair were as efficiently sequenced. The average overlapping length was 54 (min=6, max=100) and quality score per read was of 34.95, therefore excellent. Before mapping, we filtered out bad reads on three criteria: quality score < 30, 3 or more miscalled bases, low sequences.

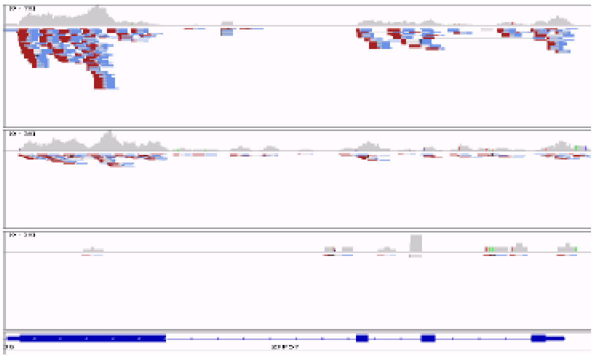


Fig. 1. IGV browser capture zooming on ZFP57. Three samples are displayed: COX at the top, an equal mix of COX and PGF in the middle, and PGF at the bottom. The number of reads is about twice higher in COX than in the mixed sample, suggestive of an allele-specific expression. *ZFP57* appears 7 times less expressed in PGF than in COX as previously reported with MHC microarrays [2].

Reads were mapped on the hg19 assembly that contains PGF sequence in the MHC region. We artificially created two other genomes with a composite chromosome 6 where we “grafted” the COX or QBL

MHC sequence. PGF samples were mapped against the 3 genomes, COX and PGF-COX mix against hg19 and the COX composite genome, QBL against hg19 and the QBL composite genome. After mapping, aligned reads are visualized in the IGV browser. Fig 1 confirms allele-specific expression for ZFP57, the top differentially expressed MHC gene. A pipeline of RNA-seq data analysis has been developed to automate the process from raw data to the quantification of genome-wide transcripts.

3 Quantification of HLA Genes Specific Expression Through the Validation of a New Algorithm for RNA-Seq Data

The genotyping of HLA genes is most challenging considering their sequence diversity, making difficult to discriminate two similar HLA alleles that differ by only one SNP. Recent studies have exploited NGS data to characterize HLA types up to the 8-digit level. Their approach mostly uses read data produced from HLA-amplified DNA and map them to HLA alleles of IMGT-HLA database (http://www.ebi.ac.uk/ipd/imgt/hla/) with Blast or Blat, allowing to genotype classical HLA genes [4,5,6]. Here, our objective is also quantitative: quantifying the overall expression of classical HLA-genes and their allele-specific expression. First, we take an input set of reads putatively mapping to HLA genes, after removal of reads mapping elsewhere in the genome. As an example for sample 1, 95,173 paired-end reads specifically mapping to HLA-genes and 3.7 millions genome-wide unmapping paired-end reads are saved as the input for the HLA-algorithm. Paired-end reads are then mapped using typical NGS mapping software, Bowtie2 or Tophat [3]. The stepping stone of the workflow not previously explored is based on specific mapping, where we can map a read to one allele, in a hash table efficiently structuring HLA alleles. This information helps localizing the other reads by reducing the number of candidate alleles. Mapping process iterates until gene level, or a 2 to 4 digit-level genotyping resolution, depending on RNA-Seq depth.

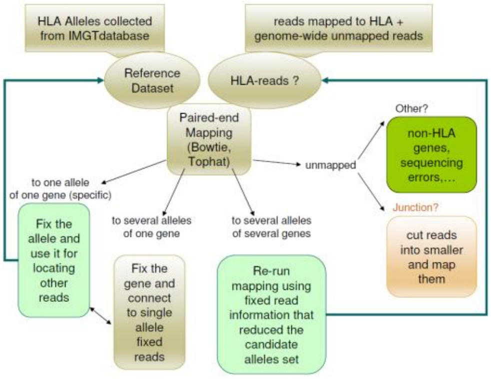

HLA Alleles collected from IMGTdatabase
reads mapped to HLA + genome-wide unmapped reads
Reference Dataset
HLA-reads ?
Paired-end Mapping (Bowtie, Tophat)
unmapped
Other?
non-HLA genes, sequencing errors,…
Junction?
cut reads into smaller and map them
to one allele of one gene (specific)
to several alleles of one gene
to several alleles of several genes
Fix the allele and use it for locating other reads
Fix the gene and connect to single allele fixed reads
Re-run mapping using fixed read information that reduced the candidate alleles set

Fig. 2. HLA Mapping Strategy

References

[1] J. Z. Levin1, M. Yassour1, X. Adiconis1, C. Nusbaum1, D. A. Thompson, N. Friedman, A. Gnirkel and A. Regev, Comprehensive comparative analysis of strand-specific RNA sequencing methods, *Nat Methods*, 7:709-715, 2010.

[2] C. Vandiedonck, M. S. Taylor, H. E. Lockstone, K. Plant, J. M. Taylor, C. Durrant, J. Broxholme, B. P. Fairfax and J. C. Knight, Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. *Genome Res*, 21: 1042-1054, 2011

[3] C. Trapnell, L. Pachter and S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25: 1105-1111, 2009

[4] T. Shiina, S. Suzuki, Y. Ozaki, H.Taira, E. Kikkawa, A. Shigenari, A.Oka, T. Umemura, S. Joshita, O. Takahashi, Y. Hayashi, M. Paumen, Y. Katsuyama, S. Mitsunaga, M.Ota, J. K. Kulski and H. Inoko, Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. *Tissue Antigens* , 80: 305-316, 2012

[5] C. Wang, S. Krishnakumar, J. Wilhelmy, F. Babrzadeh, L. Stepanyan, L. F. Su, Douglas Levinson, Marcelo A. Fernandez-Viña, Ronald W. Davis, M. M. Davis, and M. Mindrinos, High-throughput, high-fidelity HLA genotyping with deep sequencing, *Proc Natl Acad Sci U S A*, 109: 8676-8681, 2012.

[6] R.L. Warren, G. Choe, D. J Freeman, M. Castellarin, S. Munro, R. Moore, R. A. Holt, Derivation of HLA types from shotgun sequence datasets. *Genome Med*, 4: 95, 2012.

MINIA on a Raspberry Pi

Assembling a 100 Mbp Genome on a Credit Card Sized Computer

Guillaume COLLET[1], Guillaume RIZK[2], Rayan CHIKHI[2] and Dominique LAVENIER

[1] IRISA - INRIA Rennes, Campus de Beaulieu, Rennes, France
{guillaume.collet, guillaume.rizk, dominique.lavenier}@irisa.fr
[2] Department of Computer Science and Engineering, The Pennsylvania State University, USA
chikhi@psu.edu

Abstract *This work shows that the genome assembly program MINIA is able to assemble a 100 Mbp genome on a Raspberry Pi. The MINIA software was developed to drastically reduce the memory footprint needed for genome assembly, enabling human genome to be assembled on a desktop computer. Here we show that it is also able to successfully assemble a genome on a very low-end, low-power system with 512 MB RAM and a 32 GB flash drive.*

Keywords Genome Assembly, Raspberry Pi, Next-Generation Sequencing

1 Introduction

Genome assembly consists in reconstructing a sequence given only short sub-strings. It is a very difficult computational problem, as sequencers produce hundred of millions of reads shorter than a millionth the size of the genome. Furthermore, the complex repeat structure of genomes introduce ambiguities that assemblers have to solve to avoid reconstruction errors. Recently, several approches have been proposed to lower the computational requirements of targeted and whole genome assembly [1,2], down to commodity hardware. We go further and demonstrate that eukaryotic genome assembly can be performed on a Raspberry Pi, a credit-card-sized Linux computer. While this experiment is only a proof of concept and does not reflect an actual assembly pipeline, it showcases that complex *de novo* genome analyses using low computational power are now within our reach.

2 MINIA

A new data structure has been recently proposed by Chikhi and Rizk [2] to lower the memory footprint of genome assembly. The approach is based on a novel representation of de Bruijn graphs. The nodes of the graph are encoded using a Bloom filter. The edges are dynamically inferred from the nodes. A second data structure is used to remove spurious edges due to false positives of the Bloom filter. The implementation, the MINIA software, has been applied to perform the first ever assembly of a human genome on a desktop computer. Note that MINIA only produces contigs, thus does not perform all the tasks (scaffolding, gap-filling) necessary to produce a high-contiguity assembly.

3 Raspberry Pi

The Raspberry Pi is a low-cost (35 €) credit-card sized computer developped by the Raspberry Pi Foundation.[1] The aim of this foundation is to promote computer science education for all. The developpement of the Raspberry Pi was thus guided by low costs and good performances. In this work, MINIA was benchmarked on the last version of the Raspberry Pi which is based on an ARM11, 700 MHz processor with 512 MB RAM. The Linux system (RaspDebian) was hosted on a 16 GB sdcard. The benchmark files were put on a 32 GB flash drive.

[1] http://www.raspberrypi.org/about

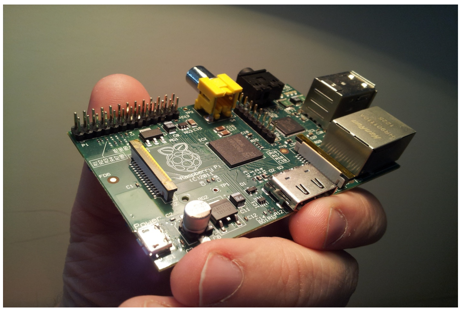


Figure 1. The Raspberry Pi is a low-cost (35€) credit-card sized computer.

4 Results

Our experiment consists in assembling the 100 Mbp genome of the nematode *C. elegans*. We used 33 million unfiltered paired-end reads of length 100 bp (SRR065390), covering the genome at about 64x. We compared the assembly results of MINIA (v1.4961) running on a Raspberry Pi, with the assembly results computed by Velvet (v1.2.08) [3] and SOAPdenovo (v2.04) [4] running on a 64 GB system equipped with a Xeon E5462 running at 2.8GHz. The QUAST software [5] was used to assess the quality of contigs generated with each method (Table TABLE 1). SOAPdenovo was run to output contigs only. The *-scaffold* option of QUAST was used to evaluate the contigs generated by Velvet. Note that Velvet and SOAPdenovo performed paired-end assembly improvement steps (scaffolding and gap-filling), that were not evaluated. To obtain a fair comparison with MINIA, we focused on contigs quality only, which is arguably the central assembly task. None of the assemblies was better over all metrics. MINIA is slightly behind SOAPdenovo in terms of N50 (5741 vs 5975) and misassemblies (12 against 7). However, MINIA used a hundred times less memory than both assemblers (0.2 GB vs 29.6 GB and 30.6 GB). This shocases that using very little memory is not incompatible with providing close to state-of-the-art contigs quality.

Method	MINIA	SOAPdenovo	Velvet
System	Raspberry Pi	64GB/Xeon E5462	64GB/Xeon E5462
CPU time (h)	18.9	6.25	13.5
Peak memory (GB)	0.2	29.6	30.6
Number of contigs (K)	29.5	29.5	28.2
Longest contig (Kbp)	75.2	90.9	62.6
Contig N50 (bp)	5741	5975	6031
Sum (Mbp)	86.4	88.3	90.4
Misassemblies	12	7	419
Genome fraction (%)	80.9	82.8	85.0
mismatches (per 100 Kbp)	3.2	0.75	25.6

Table 1. De novo *C. elegans* contigs assembled by MINIA [2], SOAPdenovo [4], and Velvet [3]. Assembly quality was evaluated using the QUAST software [5]. MINIA and Velvet were single-threaded. For SOAPdenovo, the CPU time is the sum for each thread.

References

[1] P. Peterlongo, & R. Chikhi (2012). Mapsembler, targeted and micro assembly of large NGS datasets on a desktop computer. BMC Bioinformatics, 13(1), 48. doi: 10.1186/1471-2105-13-48

[2] R. Chikhi, & G. Rizk (2012). Space-efficient and exact de Bruijn graph representation based on a Bloom Filter. Lecture Notes in Computer Science (Ed.), WABI, 7534, 236-248

[3] D. Zerbino, & E. Birney (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res., 18, 821-829

[4] R. Li, et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res., 20, 265-272

[5] G. Alexey, et al. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075

Development of a Composition Based Filter for Identification of Protein Tandem Repeats

François RICHARD[1,2] and Andrey KAJAVA[1,2]

[1] Centre de Recherche de Biochimie Macromoléculaire (CRBM), UMR5237 CNRS, 1919, Route de Mende, 34293, Montpellier, Cedex 5, France

[2] Institut de Biologie Computationnelle (IBC), 95 rue de la Galéra, 34095, Montpellier, France

francois.richard@crbm.cnrs.fr, andrey.kajava@crbm.cnrs.fr

Keywords Proteomes, tandem repeats, filter, short strings, composition.

1 Introduction

Today, the growth of the protein sequencing data significantly exceeds the growth of capacities to analyze these data. In line with the dramatic growth of this information and urgent needs in new bioinformatics tools our work deals with the development of new algorithms to better understand the sequence-structure-function relationship. A majority of protein sequences are aperiodic and usually have globular 3D structures carrying several different functions. The foremost efforts of researchers were devoted to these types of proteins and, as a result, significant progress has been made in the development of bioinformatics tools for their analysis. However, proteins also contain a large portion of periodic sequences representing arrays of repeats that are directly adjacent to each other [1], so called tandem repeats (TRs). TRs occur at least in 14% of all proteins [2]. Moreover they are found in every third human protein. Highly divergent, they range from a single amino acid repetition to domains of 100 or more repeated residues. Over the last decade, numerous studies demonstrated the fundamental functional importance of such TRs and their involvement in human diseases. A number of evidences have also been gathered about the high incidence of TRs in the sequences of virulence factors of pathogenic agents, toxins and allergens [3]. The genetic instability of these regions can allow a rapid response to environmental changes, and thus, can lead to emerging infection threats. This implies that this class of sequences may have a broader role in human diseases than was previously recognized. Thus, TR regions are abundant in proteomes and are related to major health threats of the modern society. Along this line, the discovery of these domains, understanding of their sequence–structure–function relationship and mechanisms of their evolution promise to be a fertile direction for research leading to the identification of targets for new medicines and vaccines.

2 Necessity of an Algorithm to Speed Up Identification of TR Regions

The growth of proteomic data has led to increasing efforts to develop methods for protein repeat recognition. Protein TRs are frequently not perfect, containing a number of mutations (substitutions, insertions, deletions) accumulated during evolution, and some of them cannot be easily identified. One of the most sensitive approaches for *ab initio* identification of “covert” TRs relies on Hidden Markov Model (HMM) – HMM comparisons like HHrepID [4]. However, this kind of method is relatively slow and inappropriate for automated large-scale analysis [5]. In this situation, algorithms allowing the speeding up of the TR analysis are of urgent need.

3 Development of Composition Based, Rapid Pre-filtering to Select Proteins that May Contain TRs

In line with this need, we are working on an algorithm that will pre-filter proteins with TRs. The general idea behind this project is to use, prior to the existing TR identification programs, an approach that is able to rapidly pre-select proteins with a high probability to have TRs, while discarding information about their exact location, and repeat alignments. For TRs with short repeats (from 1 to 20 residues) we apply filters that are based on an estimation of amino acid compositions in a given window. For longer TRs, we focus on the analysis of the composition of two residue strings and their order of appearance within proteins. We assume that TR regions contain a higher frequency of certain shorts strings compared to the aperiodic sequences. Therefore, we established a process of comparison of occurrence of the short strings in the neighbouring sliding windows. The score between two windows is defined as the number of common short strings between them and for which the relative order has been conserved. To improve performance of the filters we analyse short strings taking into consideration their similarity in terms of physico-chemical properties and by using different matrices of amino acid substitutions. It was also shown that the filtering can be improved by clustering the proteins using CD-HIT [6], when at least one protein from a cluster is revealed as a TR-containing protein then the other members of this cluster are considered to have TRs.

To test the performance of our filters we built two datasets. First, negative set contains randomly generated sequences that are assumed to be without TR regions generated and maintained during evolution. This set consists of groups of randomly generated sequences from 100 residues to 1700 residues increasing by a step of 50 residues. For each group, the initial sequence has been generated by MAKEPROTSEQ, a program from EMBOSS [7], using the average composition of SwissProt database and then randomize 1000 times by SHUFFLESEQ, a program from EMBOSS [7], given 1000 sequences. Second, positive set consists of TR containing protein sequences identified by HMMs and coming from the families of tetratricopeptide repeats (TPR), leucine-rich repeats (LRR) and WD repeats. Those repeats cover α-helices, α/β-structures and β-structures and are known to be among the most degenerated ones and therefore, the hardest to find by *ab initio* methods.

4 Current Status of Work and Future Plans

Our preliminary tests allow to identify 98% TR-containing proteins from the positive set and to discard up to 25% of the proteins from the negative set. This unveils a new approach to quickly reduce the initial amount of initial proteins prior to TR identification. We are working now on the optimisation of the algorithms and their parameters.

Acknowledgements

Funding: Ministère de l'Enseignement supérieur et de la Recherche (MESR) grant to F.R.

References

[1] J. Heringa, Detection of internal repeats: how common are they? Curr. Opin. Struct. Biol. 8:338–345, 1998.

[2] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, Detecting Protein Function and Protein-protein Interactions from Genome Sequences. Science, 285:751–753, 1999.

[3] A. V. Kajava, J. M. Squire, and D. A. D. Parry, Beta-structures in Fibrous Proteins. Advances in Protein Chemistry 73:1–15, 2006.

[4] A. Biegert and J. Söding, De novo identification of highly diverged protein repeats by probabilistic consistency, Bioinformatics, 24:807–814, 2008.

[5] A. V. Kajava, Tandem Repeats in Proteins: From Sequence to Structure. *Journal of Structural Biology* 179:279–288, 2012.

[6] L. Weizhong, and A. Godzik, Cd-hit: a Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. Bioinformatics 22:16581659, 2006.

[7] P. Rice, I. Longden, and A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite, *Trends Genet.* 16:276277, 2000.

Evolutionary dynamics and unexpected bacterial content in closely related genomes of the model pathogen fungus *Magnaporthe*

Ludovic MALLET[1,2], Cyprien GUÉRIN[1], Gabriela AGUILETA[3], Joelle AMSELEM[4,6], Enrique ORTEGA-ABBOUD[5], Didier THARREAU[5], Marc-Henri LEBRUN[6], Elisabeth FOURNIER[7] and Hélène CHIAPELLO[1,8]

[1] INRA, UR MIG, Domaine de Vilvert, 78352 Jouy-en-Josas, France
{ludovic.mallet, cyprien.guerin}@jouy.inra.fr

[2] Evolutionary Bioinformatics Group, Institute for Evolution and Biodiversity, Westfälische Wilhelms-Universität, Hüfferstrasse 1, D-48149 Münster, Deutschland
ludovic.mallet@uni-muenster.de

[3]Comparative Genomics Group, CRG-Center for Genomic Regulation, Doctor Aiguader 88, lieu, 08003 Barcelona, Spain
gabriela.aguileta@crg.eu

[4]URGI, INRA, Route de St Cyr, 78026 Versailles, France
joelle.amselem@versailles.inra.fr

[5]UMR BGPI, CIRAD, TA 54K, 34398 Montpellier, France
{enrique.ortega-abboud, didier.tharreau}@cirad.fr

[6]UMR BIOGER, INRA, 78850 Thiverval-Grignon, France
marc-henri.lebrun@versailles.inra.fr

[7]UMR BGPI, INRA, TA 54K, 34398 Montpellier, France
elisabeth.fournier@supagro.inra.fr

[8]UR MIAT, INRA, chemin de Borde-Rouge, 31326,Castanet-Tolosan, France
helene.chiapello@toulouse.inra.fr

Abstract. We have compared and analyzed a dataset of ten closely related genomes of the *M. Oryzae/grisae* species complex, a model pathogen fungus infecting rice and other Poaceaes. We detected unexpected large supplementary genomic regions potentially issued from an unknown bacterial strain of the *Burkholderia* genus in four of our genomes. We used a parametric method based on genomic signature to accurately quantify, characterize and filter these regions in all the affected genome scaffolds. Systematic detection of potential horizontal transfers and transposable elements was also carried out in the 9 newly sequenced genomes. Finally, by using predicted *M.grisae/oryzae* orthologs families processed by a 2-step Bayesian analysis, we were able to infer the phylogenetic reference genealogy of the *M. oryzae/grisae* complex. We will present preliminary results regarding the comparison of TE distribution in *M. oryzae/grisea* genomes taking into account the reference genealogy of the strains.

Keywords Genome evolution, Phylogenomics, Horizontal Transfers, Transposable Elements.

CSP4CSP: a Constraint Programming Approach for Scaffolding

Annie CHATEAU[1,2,3], Nicolas BRIOT[2], Simon GIVAUDAND[2], Francisco RODRIGUEZ GOMEZ[2] and Olivier SANS[2]

[1] LIRMM, UMR 5506, 161 rue Ada 34095 Montpellier Cedex 5 - France, chateau@lirmm.fr
[2] Université Montpellier 2, 2 Place Eugène Bataillon 34095 Montpellier Cedex 5 - France
[3] Institut de Biologie Computationnelle (IBC), 95 rue de la Galéra, 34095 Montpellier, France

Keywords Scaffolding, Constraints Solver, NGS

1 Introduction

We are presenting in this contribution a work in progress concerning the genome scaffolding problem. Next Generation Sequencing provides more and more data. The corresponding issues like genome assembly or how to extract genomic information from those data have been widely discussed. We focus here on the scaffolding step, that is, once the *de novo* assembly of a genome has been performed, leading to a set of contigs, compute an ordering and orientations of the contigs in the best possible way. Contig scaffolding has been stated as a NP-complete problem, whose solutions can be approximated by a greedy approach [1]. Mostly used scaffolders, like Opera [2], Sopra [3], GRASS [4], use different kinds of heuristic or statistical approaches. In this work, we propose an original way to consider the Contig scaffolding Problem, which could eventually be adapted to various sources of information on the orientation and the order of the contigs. We adopt here a black-box approach, using the Constraint Programming paradigm. We encode the Contig Scaffolding Problem into a Constraint Solving Problem (CSP). In this short paper we present our modelization, and how we plan to test this approach.

2 Contig Scaffolding Problem

The scaffolding problem is defined as follows: given a set of contigs $\mathcal{C} = \{C_1, \ldots, C_n\}$, and a set of paired-end reads, we want to infer an order on the contigs, and the orientation of the contigs, that is the most coherent with the pairs. We consider that the reads have already properly been mapped to the contigs and we focus on pairs whose elements are mapped on distinct contigs. Given a fixed orientation of contigs, there are four ways, called stories, in which a pair can map on two distinct contigs, depending on the reads orientations relative to the contigs orientations (see Fig. 1).

Story A (resp. B) tells that contigs C_1 (resp. C_2) preceeds contig C_2 (resp. C_1) with their original orientations. Story C (resp. D) tells that contig C_1 (resp. C_2), with its alternative orientation, preceeds contig C_2 (resp. C_1).

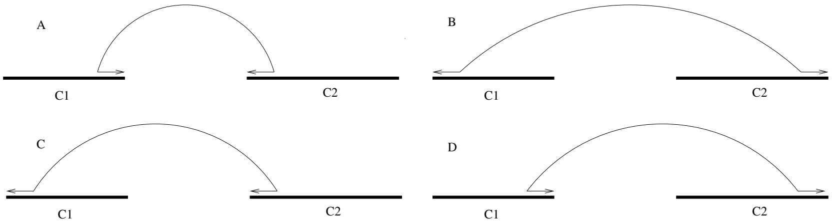


Figure 1. The four possible ways of linking two contigs.

Several pairs could tell the same stories. In this case, we consider that we can bundle them and attribute a weight to the story equal to the number of stories that have been bundled. The data of the contigs and the weighted stories is called the scaffold graph. We denote by m the number of distinct stories, meaning the number of edges in the scaffold graph.

3 CSP4CSP: Constraint Solving Problem

The CSP modelization is the following: each contig C_i gives two variables C_i^b (begin) and C_i^e (end) whose domain is $\{1, \ldots, 2n\}$; each story gives a variable H_j whose domain is $\{0, 1\}$. A value 0 means that the story is not chosen, 1 means that the story is chosen. We put the following constraints on these variables:

- $Alldiff(C_i^\alpha, i \in \{1, \ldots, n\}, \alpha \in \{b, e\})$ stating that the contig variables take distinct values;
- $\forall i \in \{1, \ldots, n\}$ $|C_i^e - C_i^b| = 1$, meaning that both extremities of C_i are consecutive;
- $\forall k \in \{1, \ldots, m\}$, if $C_{i_k}^{\alpha_k}$ and $C_{j_k}^{\beta_k}$ are the contig extremities in the story H_k: $H_k(1 - |C_{i_k}^{\alpha_k} - C_{j_k}^{\beta_k}|) = 0$, meaning that if the story H_k is chosen, then the concerned contig extremities are consecutive.

We use a max CSP solver, which maximizes the following objective function: $w = \sum_{k=1}^{m} H_k.w_k$, where w_k is the weight of the story H_k. Notice that w represents the total number of pairs supporting the chosen solution.

4 Implementation

A JAVA implementation is being developped, which extracts the stories from the mapping data (SAM file) and transform the scaffold graph into CSP instances. The function that bundles the stories is a parameter, so we could study its influence on the accuracy of the result; A solver output parser, and several tools to measure the quality of the produced scaffold are on their way to complete this tool. We use Numberjack, a modelling package in Python for constraint programming, with the CSP solver Mistral [5]. An example on a toy graph can be found in Fig. 2.

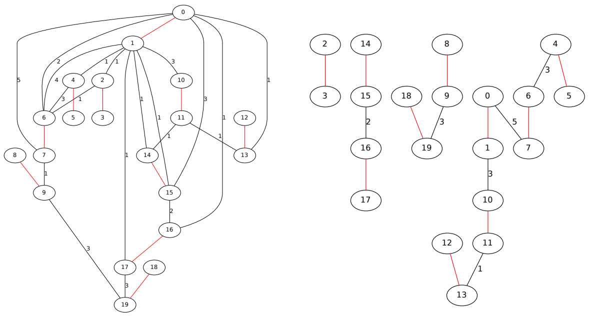


Figure 2. The original graph (left) and its optimal linearization (right). Contigs are unlabeled edges.

5 Experiments

In a first phase we use simulated data, producing original sequences and paired-end reads. Several aspects will be studied: influence of the read cover, choice of the assembly software, influence of the mapping parameters, limitation due to the size of the genome, time performances. To compare the results, we focus on two kinds of quality measures: the usual N50 criteria, and since we know the original sequence, we can map the contigs on the original sequence, and infer the "right order", to be compared to the one produced by our solver. In a second phase, the method will be used with optimized parameters on real data.

Acknowledgements

We would like to thank Remi Coletta[1,2], our "Constraint Master" for his useful advices. This work is supported by the Institut de Biologie Computationnelle (IBC) and by the Mastodons challenge.

References

[1] D. H. Huson, K. Reinert, and E. W. Myers, The greedy path-merging algorithm for contig scaffolding, *Journal of the ACM* vol 49, 5:603-615, 2002.

[2] S. Gao, W.-K. Sung, and N. Nagarajan, Opera: Reconstructing Optimal Genomic Scaffolds with High-Throughput Paired-End Sequences. *Journal of Computational Biology*, vol. 18, 11:1681–1691, 2011.

[3] A. Dayarian, T. P. Michael, A. M. Sengupta, SOPRA: Scaffolding algorithm for paired reads via statistical optimization, in Dayarian et al. (eds), *BMC Bioinformatics*, 11:345, 2010.

[4] A. Gritsenko, J. Nijkamp, M. Reinders and D. de Ridder, GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies, *Bioinformatics*, 2012.

[5] E. Hebrard, E. O'Mahony and B. O'Sullivan, Constraint Programming and Combinatorial Optimisation in Numberjack, in A. Lodi, M. Milano and P. Toth (eds), *Proceedings of the 7th International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CPAIOR-10)*, LNCS, 6140:181–185 , Springer-Verlag 2010.

EMME : an Experimental Metadata Management Environment

Cyril MONJEAUD[1], Yvan LE BRAS[1] and Olivier COLLIN[1]

[1] INRIA/IRISA Campus de Beaulieu, 35042, Rennes, Cedex, France
{Cyril.Monjeaud, Yvan.le_bras, olivier.collin}@irisa.fr

Abstract *The EMME project (Experimental Metadata Management Environment) carried out by the GenOuest bioinformatic facility, aims at developing a metadata tracking platform for the technical facilities in the Western France. The EMME environment is based on ISA tools in association with some analysis tools. A web portal is designed and provides the ISA Software Suite (ISA-Tab creation), a Galaxy instance (data analysis), two FTP servers (storage/sharing) and an OMERO instance (images visualization).*

Keywords Metadata, web portal, ISA-Tab, data analysis, life science

EMME : un EnvironneMent de gestion des Métadonnées Expérimentales

Résumé *Le projet EMME (EnvironneMent pour la gestion des Métadonnées Expérimentales), développé par la plate-forme bio-informatique GenOuest, a pour objectif la création d'un environnement proposant une gestion des métadonnées expérimentales. Ce projet cible les différentes plate-formes technologiques de l'Ouest de la France. L'environnement EMME est basé sur les outils ISA (Investigation, Study, Assay). Un portail web a été réalisé et propose différents outils allant de la suite logicielle ISA pour la gestion des métadonnées aux services de stockage et d'analyse tels qu' Ajaxplorer, Galaxy ou OMERO.*

Mots-clés *Métadonnées, portail web, ISA-Tab, analyse de donnée, Sciences de la vie*

1 Contexte : les métadonnées expérimentales

Aujourd'hui, les nouvelles technologies produisent une grande quantité de données. Chaque donnée est amenée à être analysée plusieurs fois de différentes façon ou à être associée à d'autre informations. Afin d'optimiser le travail des équipes, les données doivent être standardisées, facilement identifiables et facilement accessibles. Ceci est possible en utilisant les métadonnées expérimentales, qui décrivent comment les données ont été produites ainsi que leur localisation, et en mettant en place une infrastructure et un ensemble d'outils permettant de les capturer. C'est ce que propose le projet EMME.

Ce projet a pour vocation de faciliter l'intégration et l'échange de données au sein des différentes entités de Biogenouest. Le portail propose l'accès à des outils ISA (Investigation, Study, Assay) permettant une prise en charge des métadonnées expérimentales via le format ISA-Tab [1] et leur dépôt dans le modèle « BioInvestigation Index » (BioInvIndex ou BII).

Le point fort de ce système est la possibilité de gérer des expériences multi 'omique' (protéomique, génomique, métabolomique, etc). Ainsi, les informations et les données de chaque expérience réalisée sont plus facilement accessibles et pérennisées.

2 Résultat : Le portail EMME

Le portail EMME est accessible à l'adresse http://emme.genouest.org. Les outils proposés par ce portail sont brièvement décrits ci-dessous.

2.1 Les services dédiés aux métadonnées

Une équipe, basée à l'Oxford e-research center, propose des outils permettant la prise en charge des métadonnées ainsi que leur stockage. Débutant en 2008, cette équipe a mis en place le format standard ISA-Tab et a développé un modèle pour le stockage de ces métadonnées ainsi qu'une suite logicielle.

Le **format** général (ISA-Tab) est un format tabulé simple dédié aux biologistes et aux bio-informaticiens. Le but étant de faciliter la gestion locale des métadonnées expérimentales (caractéristiques des échantillons, technologies et types de mesure, relations échantillon-donnée) pour les études utilisant une ou plusieurs technologies.

Pour finir, l'application web BioInvIndex (BII) accessible via le portail EMME est basé sur le **modèle** de stockage des expériences ISA-Tab mentionné plus tôt. Cette interface met à disposition toutes les expériences ISA-Tab déposées dans la base de données (maintenue par l'administrateur). L'utilisateur peut ainsi rechercher une étude/expérience et consulter les détails associés. Ce dernier peut également télécharger le fichier ISA-Tab correspondant à l'expérience consultée ainsi que les fichiers de données dans le cas où ils sont disponibles. Une étude peut ainsi être rendue privée et accessible uniquement à une ou à un ensemble de personnes.

La **suite logicielle** ISA Software Suite composée des outils ISACreator et ISACreatorConfigurator permet de capturer les métadonnées expérimentales et de créer une archive ISA contenant les fichiers ISA-Tab. Cette dernière sera prise en charge par nos outils d'analyse afin de rapatrier les données et métadonnées d'une expérience. L'utilisation de cette suite logicielle peut se faire via le portail EMME sans aucune authentification.

2.2 Les services dédiés aux données

2.2.1 Le logiciel Ajaxplorer

Le logiciel Ajaxplorer[1] est un gestionnaire de fichiers en ligne gratuit et open-source (sous licence AGPL). Une instance d'Ajaxplorer a été mise en place dans le cadre du projet EMME et est accessible sur le portail.

Transformé en client FTP, ce logiciel donne accès aux deux serveurs détaillé ci-dessous. Seuls les utilisateurs ayant un compte GenOuest sont autorisés à accéder à ce service (inscription gratuite sur le site de la plate-forme[2]). Une fois connecté sur Ajaxplorer, ces deux serveurs seront accessibles en parallèle ; un transfert de fichier entre les deux emplacements est donc facilement réalisable. La possibilité de générer un lien public par fichier est également un point crucial pour le partage.

2.2.2 Le serveur FTP EMME

Ce serveur FTP propose un espace de stockage pour les personnes utilisant les services dédiés aux métadonnées (décrits ci-dessus). L'utilisateur sauvegarde les données issues des expériences sur ce serveur et cette localisation est renseignée dans le fichier ISA-Tab correspondant. Ce dernier point est essentiel pour la pérennité des données.

2.2.3 Les serveur FTP Galaxy

Contrairement au serveur FTP précédant, ce dernier n'est pas directement dédié au stockage de fichiers. En effet, il a pour but de rendre disponible un fichier à l'instance Galaxy de GenOuest de façon temporaire. Une fois importé dans cette dernière, le fichier sera automatiquement supprimé du serveur FTP.

[1] http://ajaxplorer.info

[2] http://genoweb1.irisa.fr/AppliESIB/access/access-en.php

2.3 Les services dédiés aux analyses

2.3.1 Galaxy

Galaxy [2, 3, 4] est une plate-forme web dédiée aux analyses biologiques/bio-informatiques. Les utilisateurs sans notions en programmation peuvent facilement spécifier divers paramètres afin d'exécuter des outils ou des workflows. Cette plate-forme d'analyse permet à l'utilisateur de répéter et de partager / publier ses analyses via le web.

Une instance de Galaxy a été déployée par GenOuest afin de mettre à disposition sa puissance de calcul sans utiliser de lignes de commandes. L'accès à ce service nécessite également un compte sur la plate-forme GenOuest. Le serveur FTP Galaxy présenté ci-dessus est directement couplé à cette instance.

Des outils, dont le point d'entrée est une ISArchive, ont été mis en place pour interagir avec des fichiers ISA-Tab. Galaxy propose d'une part un outil capable d'extraire les fichiers ISA-Tab de l'archive, mais également d'importer automatiquement dans l'historique courant tous les jeux de données publics mentionnés dans ces derniers. Pour finir, un outil à été développé pour ajouter directement son ISArchive dans le BioInvIndex de GenOuest.

2.3.2 OMERO

OMERO [5] est un logiciel client-serveur pour la visualisation, la gestion et l'analyse des images issues de la microscopie en biologie. Ce logiciel dispose d'outils capables d'importer des images en les organisant au sein de projets.

De plus, OMERO permet un visionnage des images et leur analyse via des scripts. Pour finir, l'export d'images pour l'analyse ou la publication est également réalisable.

Acknowledgements

This work was supported by an IbiSa grant 2011.

References

[1] Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S, Hide W, Hofmann O, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*. 2010.

[2] Goecks J, Nekrutenko A, Taylor J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010. 25;11(8):R86.

[3] Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*. 2010. Chapter 19:Unit 19.10.1-21.

[4] Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*. 2005. 15(10):1451-5.

[5] Allan C., Burel JM, Moore J, Blackburn C, Linkert M, Loynton S, MacDonald D, Moore W, Neves C, Patterson A, Porter M, Tarkowska A, Loranger B, Avondo J, Lagerstedt I, Lianas L, Leo S, Hands K, Hay R, Patwardhan A, Best C, Kleywegt G, Zanetti G, Swedlow J. OMERO: flexible, model-driven data management for experimental biology. *Nature Methods 9*, 245–253.

CIRCUS: utility package for CIRCos USage

Delphine NAQUIN[1], Yves D'AUBENTON-CARAFA[1], Claude THERMES[2] and Maud SILVAIN[2]

[1] Plateforme IMAGIF, FRC3115 CNRS, avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France
{delphine.naquin, daubenton}@cgm.cnrs-gif.fr

[2] Centre de Génétique Moléculaire, UPR3404 CNRS, avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France
(maud.silvain, claude.thermes)@cgm.cnrs-gif.fr

Abstract *Detection of genomic large modifications, such as large indels, duplications or translocations, is a hard task. Recently several tools have been developed to analyze NGS data but result files are difficult to interpret without an additional visualization step. Circos[1] is a powerful visualization software designed for this purpose, but it requires configuration files that need tedious devising. To overcome this difficulty, we developed R scripts called CIRCUS that manage both data and configuration files to produce easily and quickly relevant graphs. Many options are available around a fixed canvas to deal with heterogeneous data, genomes with multiple chromosomes and multi-scale analyses.*

Keywords Circos, variant genome structure, visualization

1 Introduction

Large-scale genomic variations are now usually detected by paired-end or mate pair sequencing. The two reads of a fragment that map either at abnormal positions on a chromosome, or on two different chromosomes, may point out the border of a genome variation. A list of these borders is difficult to analyze since one variation involves often two borders that can be localized at remote positions. Two visualization softwares[1,2] have been developed, where each event can be displayed as a link between positions on a circular ideogram. The most used, Circos, is very flexible but it requires programming skills and tedious devising. To circumvent this difficulty, the variant detection program SVDetect[3] offers a script to convert its results into a format readable by Circos and proposes a set of configuration files. However, the drawing is freezed with no possibilities for example to zoom on regions of interest. In this context, we present a set of R scripts called CIRCUS that uses output files of several variant structure detection tools, to fill all necessary files for Circos execution with many options to customize a quick and flexible image production.

2 Architecture and workflow

Besides the positions and the significance of the event borders, the supplementary data suitable for genomic variations analysis are the local coverage in reads, the CNV inference and the genomic annotations. CIRCUS allows to display all these features on a fixed framework which consists in 3 optional concentric rings containing from inner to outer the CNV inference in colored windows, the read coverage in histogram style and the genomic annotations in colored boxes (see figure 1). Around the center of the image, links can be painted according to their significance. The rings can be divided into two main parts. The first one contains one or two regions called view(s) within a chromosome of interest. The second is designed to display a set of entire chromosomes as well as an optional pseudo-chromosome (referenced later as NM for No Match chromosome) which can sign an integration site of a foreign DNA fragment. Only links with at least one foot in the view(s) will be displayed.

The core of CIRCUS is a R function which allows the user to specify the dimensions of the components and the format of the picture, the type of features to be displayed and the coordinates of the region(s) to be analyzed. According to the input parameters, this function filters the links and creates the configuration files required by Circos. To feed the core, five functions are proposed that handle the data for genomic annotation, reads coverage, NM links, intra-genomic links and CNV status. They have to call specific scripts to convert the data issued from different prediction tools. Currently, reads coverage and NM links are computed from sam files with samtools/bedtools packages and a Python script. Genomic annotations from GFF as well as intra-genomic links issued from SVDetect are treated with wrapped R functions. The addition of new converters is easy and development of several new ones is in progress. The core and peripheral functions have many input arguments to allow flexibility, but to simplify the filling of these arguments, almost all have default values. CIRCUS is available on request at sequencage.bioinfo.imagif@cgm.cnrs-gif.fr.

CIRCUS has been devised to allow biologists without strong programming skills to visualize quickly and easily complex genomic variations. A typical analysis may include an iterative view of the links of each chromosome against all others followed by zooms on regions of interest. The borders of each event can thus be precisely delineated whatever the size of the corresponding DNA fragment. Localization of insertion of foreign sequences such as viruses can be detected by the links toward the NM pseudo-chromosome.

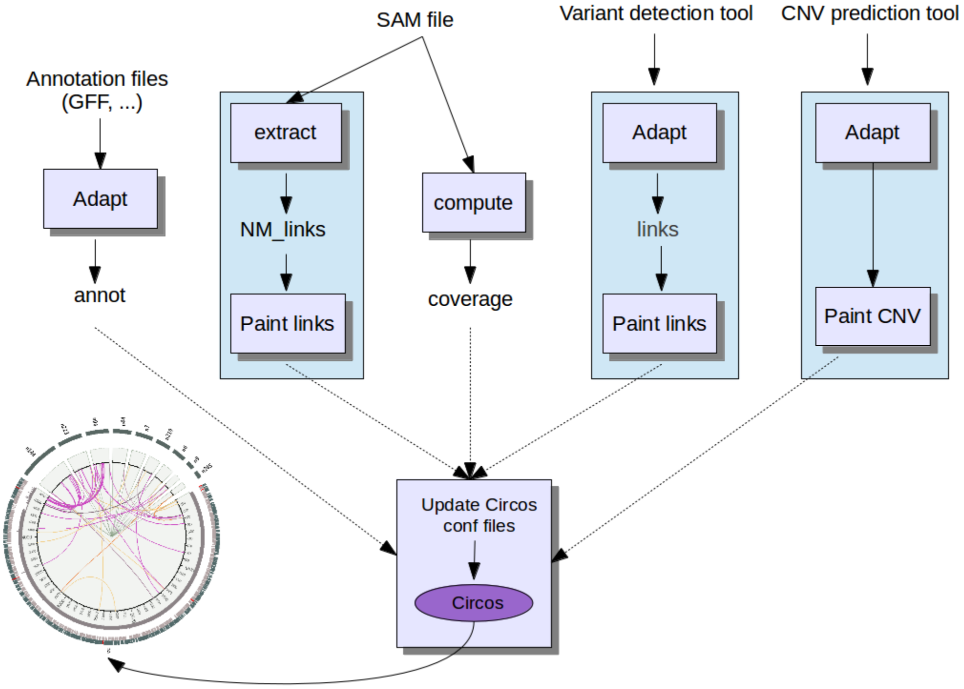


Figure 1. CIRCUS and satellite scripts.

Acknowledgements

The authors thank Mireille Ansaldi for providing data used for the illustration.

The work is supported by the Centre National de la Recherche Scientifique and the IMAGIF sequencing platform.

References

[1] Krzywinski, M. *et al.* Circos: an Information Aesthetic for Comparative Genomics. *Genome Res.*, 19:1639-1645, 2009

[2] Sven Ekdahl and Erick L. L. Sonnhammer. ChromoWheel: a new spin on eukaryotic chromosome visualization. Bioinformatics (2004) 20 (4): 576-577. doi: 10.1093/bioinformatics/btg448

[3] Bruno Zeitouni, Valentina Boeva, Isabelle Janoueix-Lerosey, Sophie Loeillet, Patricia Legoix-ne, Alain Nicolas, Olivier Delattre, Emmanuel Barillot. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, 26: 1895-1896, 2010

DEA: a parallelized pipeline for differential expression analysis

Jonathan Kreplak[1*], Yufei Luo[2*], Marie-Agnès Dillies[3*], Joelle Amselem[1], Olivier Inizan[1], Matthias Zytnicki[1] and Delphine Steinbach[1]

[1] Unité de Recherche en Génomique-Info, Platform URGI; INRA, Route de Saint Cyr, 78000, Versailles, France
{jonathan.kreplak, matthias.zytnicki, olivier.inizan, joelle.amselem delphine.steinbach}@versailles.inra.fr

[2] Plate-forme Bioanalyse Génomique, Institut Pasteur, 28 rue du Dr Roux, 75724 Paris, France
yufei.luo@pasteur.fr

[3] Plate-Forme Transcriptome et Epigénome, Génopole, Institut Pasteur, 28 rue du Dr Roux, 75724 Paris, France
marie-agnes.dillies@pasteur.fr

Abstract *Next Generation Sequencing is used as a leverage to solve new problems in transcriptomics. Differential expression analysis consists in measuring the activity of genes for an organism in different conditions and provides the genes which are over or under-expressed. We developed, under the platform Galaxy, a parallelized pipeline for differential gene expression adapted to RNA-seq data size, using community experience. The pipeline is aimed to be fast and easy to use for biologists.*

Keywords RNA-seq, Galaxy, Differential gene expression

1 Introduction

RNA-seq is a powerful technology to obtain NGS data on differential expression. We developed a pipeline, named DEA for Differential Expression Analysis. It performs the whole analysis, from the raw FastQ files to the list of differentially expressed genes, including several graphs for the visualization and control of the gene expression distribution. The main steps of DEA pipeline are: (i) mapping of reads on a reference genome using TopHat v1.4.1 [1], (ii) counting of reads mapped by gene using an S-MART tool [2], (iii) normalization and differential expression calling are performed by DESeq [3].

The first version of DEA is a result of a strong collaboration between computer scientists, bioinformaticians and statisticians. This collaboration led us to choose relevant tools and parameters, especially for DESeq [4]. Galaxy [5] is an open platform that facilitates collaboration since it embraces notions like accessibility, reproducibility and transparency. For instance, it provides an easy-to-use web-interface for biologists to launch the whole pipeline in one click.

From a computational point of view, chaining these tools is necessary to design and validate the steps but not sufficient for scalability analysis. To speed up computation, we decided to parallelize the mapping and the counting steps.

2 Results

To date, DEA has been successfully used on Maize and Arabidopsis thaliana data. It has been transferred to the Galaxy Tool Shed and is available to the international community(http://toolshed.g2.bx.psu.edu/). At URGI, we already faced parallelization in other contexts [6] and implemented an API working on SGE to address this question. Basically, API works in three steps: (i) input data are split into homogeneous batches. (ii) Each batch is processed individually. (iii) The result files of those processes are joined into a single output.

Mapping and counting steps of DEA were parallelized using this paradigm. For the mapping step, FastQ files are first cut into batches, themselves mapped on genome with TopHat. BAM files are joined using SAMtools. For the counting step, after a BAM-to-SAM conversion, SAM output files are also cut into batches. For each batch a script counts the reads matching a reference annotation. Results are merged in tabular files for DESeq analysis.

*These authors contributed equally to this work.

In order to validate the parallel implementation of the pipeline, we compared results obtained on an *A. thaliana* dataset [7] between the linear and parallel implementations.

Our first results showed differences between the two implementations related to the parallelization of TopHat. Indeed, 79 reads, from a total of 775,313, were not mapped on the genome. We therefore investigated why this reads linked to 18 different transcript with an intron were missing in our parallel implementation.

In a first step, TopHat aligns the reads on the genome. Putative *de novo* exons are formed if a sufficient number of reads co-localize. These new exons are linked with flanking exons to predict putative transcripts. In a second step, reads are aligned to the sequences of these putative transcripts. Since in the parallelized version, reads mapping the same exon do not necessarily co-occur in the same batch, the quorum of reads to form a new exon is not always achieved, and the potential *de novo* transcripts are missing. As a consequence, read mapping to these exons junction would not align.

Based on this result, we configured DEA pipeline so that it only uses known exons at this step, and we retained reads that only mapped on the annotation. With this scenario, linear and parallel implementations gave the same results.

Linear and parallelized DEA were both run on a 2.66 GHz dual core virtual machine with 2 GB RAM and twenty slots of our cluster. The linear version took 90 min to analyze 758,4 Mo of reads. These reads were produced by two experiments performed on two and three replicates respectively. The parallel version took 23 min to treat the same data.

3 Conclusion

We have presented a fast and easy-to-use parallelized pipeline for differential gene expression. On the first hand, our results show that based on specific parameters, TopHat can be parallelized to resolve quantitative transcriptomic problems. On the other hand, it should not be used for qualitative analysis as it could not detect all new variants. This parallel version of the pipeline will be available soon in our Galaxy instance and in the Galaxy Tool Shed. We also plan in a near future to improve the pipeline by integrating GSNAP as an alternative to TopHat [8].

Acknowledgements

This work was funded by GIS IBISA and was developed in the frame of a project managed in APLIBIO, the French network for bioinformatics platform, Ile de France part and the IFB project.

References

[1] Trapnell C, Pachter L, Salzberg SL, Tophat: discovering splice junctions with RNA-seq, *Bioinformatics*. 2009;25(9):1105-11

[2] Zytnicki M, Quesneville H, S-MART, a software toolbox to aid RNA-Seq data analysis, *PLoS One.* 2011;6(10):e25988

[3] Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11:R106

[4] Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F; on behalf of The French StatOmique Consortium, A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, *Brief Bioinform* (2012)

[5] Goecks J, Nekrutenko A, Taylor J, Team Galaxy. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*.2010.11:R86.

[6] Flutre T, Duprat E, Feuillet C, Quesneville H, Considering Transposable Element Diversification in *De Novo* Annotation Approaches. *PLoS ONE.*2011;6(1): e16526.

[7] Stroud H, Hale CJ, Feng S, Caro E, Jacob Y, et al. DNA Methyltransferases Are Required to Induce Heterochromatic Re-Replication in Arabidopsis. *PLoS Genet.*2012;8(7): e1002808.

[8] Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*.2010;26(7):873–881.

A visualization approach to analyse bacterial sRNA-mediated regulatory networks

Jonathan Dubois[1], Amine Ghozlane[1], Patricia Thébault[1,2] and Isabelle Dutour[1] and Romain Bourqui[1]

[1] Université Bordeaux 1, LaBRI, UMR5800 CNRS . 51, cours de la Libération, 33405 Talence F-33405 France
{dubois,ghozlane,thebault,dutour,bourqui}@labri.fr
[2] Laboratoire CBiB - Université Victor Segalen Bordeaux 2 146, rue Léo Saignat- 33076 Bordeaux France

Abstract *Many recent reviews show a wide spectrum of regulatory functions in bacteria where RNAs are involved. Focusing on the sRNAs that interact with mRNAs, the in silico prediction of interactions is still challenging as one of the main difficulty relies on the large proportion of false predictions. Therefore, novel strategies have to be proposed to improve the specificity of computational predictions before selecting a list of prioritized candidates for the experimental validation stage. Based on a network modeling approach, we present* rNAV*, a visualization software designed to help the identification of pertinent and reasonable sRNA-mRNA pair candidates from a list of thousand target predictions. This software has been applied to the analysis of the sRNA-mediated network of Escherichia coli.*

Keywords small bacterial RNA, regulatory network, visualization

1 Biological context

For the last decade, an increasing attention has been given to the bacterial small noncoding RNAs (*sRNA*s) with new evidences for the wide range of functional roles presented by these molecules [1]. Moreover with the extraordinary increase in sequencing capacity through new sequencing technologies, many prokaryotic transcriptomes (*e.g.* Escherichia coli...) have been explored and have revealed the existence of a plethora of small regulatory RNAs in bacteria. The identification of their mRNA targets is crucial to identify their functional activities and remains challenging with the need of a better understanding of the topological and biological constraints behind the formation of sRNA-mRNA duplexes [2,3,4]. Even with the most sophisticated bioinformatics prediction target tool, a main difficulty lies on the large proportion of false predictions. Therefore, novel strategies have to be carried out to improve the computational predictions by improving their specificity, before proposing new candidates for an experimental validation stage. To address this issue and reduce the number of *sRNA-mRNA* interactions to inspect, we propose a new visualization approach that exploits biological knowledges through appropriate filters.

2 rNAV, a visualizer software dedicated to sRNA-mediated regulatory networks

We present *rNAV*, a visualization software designed to help bioinformaticians and biologists for the identification, from list of thousands of predictions, of pertinent and reasonable sRNA gene candidates. First, the relationships between sRNAs and mRNAs are predicted and labeled by using a two step strategy combining the two well known softwares, IntaRNA [5] and DAVID [6]. While IntaRNA supports the detection of putative *sRNA-mRNA* interactions, DAVID performs gene annotation enrichments using statistical approaches according to multi-purpose biomolecular databases.

Second, the regulatory network is visually represented in *rNAV* to exploit the *mRNA*-target information to filter out the false-positive predictions. The expert selection is guided by exploiting (i) their functional activities when known, (ii) the putative conserved region in multi-targeting *sRNA*s, (iii) the exploitation of the deduced constraints from the neighborhood in the newtork. As a very brief description, *rNAV* provides a list of algorithms performing both selection and filtration. Moreover the softwre also provides a dedicated hierarchical drawing algorithm and a basic circular drawing one. In addition, it also supports the clustering in sub-graphs according to the conservative features of multi target sRNAs. *rNAV* has been developed in C++ with the Tulip framework [7] and takes full advantage of the the Tulip plugins management system that provides a huge range of algorithmq to improve and guide the network exploration (for instance, color mapping algorithms, layout algorithms, sub-network selection...).

*r*NAV has been applied to the *Escherichia coli K12* bacteria where more than 80 sRNAs have been experimentally validated so far.The regulatory network has been generated with 60346 edges modeling the predictive interactions between the 85 sRNA and 4142 mRNA. The specificity given by intaRNA has been improved by using our visualization approach and has helped to guide the data analysis of two known sRNAs. For both of them, we proposed new mRNA targets and sRNA regulators pairs for experimental validation in accordance with know biological features.

Acknowledgements

This work was supported by EVIDEN of the ANR-JCJC-2012.

References

[1] G. Storz, J. Vogel and K.M. Wassarman, Regulation by Small RNAs in Bacteria: Expanding Frontiers. *Mol Cell*, 43:880-891, 2011.

[2] A.S Richter,and R. Backofen, Accessibility and conservation: general features of bacterial small RNA-mRNA interactions? *RNA Biol*, 9:954-965, 2012.

[3] C. Beisel and G Storz, Base pairing small RNAs and their roles in global regulatory networks. *FEMS Microbiol Rev, Cell Biology and Metabolism Program*, 34:866-882.

[4] A. Peer and H Margalit, Accessibility and Evolutionary Conservation Mark Bacterial Small-RNA Target-Binding Regions. *J Bacteriol*, 193:1690-1701, 2011.

[5] A. Busch, A.S. Richter and R. Backofen, IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions, *Bioinformatics*, 24:2849-2856, 2008.

[6] D.W Huang, B.T. Sherman and R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc*, 4:44-57, 2009.

[7] D. Auber, P. Mutzel and M. Junger, Chapter Tulip- A Huge Graph Visualization Framework *Graph Drawing Software, Springer-Verlag*, 2003

Modeling the pocket-ligand pairs space in a perspective of a better characterizing of the protein-ligand interactions

Leslie REGAD[1], Delphine FLATTERS[1], Colette GENEIX[1] and Anne-Claude CAMPROUX[1]

[1] MTi Université Paris Diderot, UMR-S973 INSERM, 35 rue Hélène Brion, 75205, Paris, Cedex 13, France

{leslie.regad, delphine.flatters, colette.geneix, anne-claude.camproux}@univ-paris-diderot.fr

Keywords : Proteochemometric, protein-ligand interactions, multivariate analysis data, pharmacophores.

Pockets are today at the cornerstones of modern drug discovery projects and at the crossroad of several research fields, from structural biology to mathematical modeling. Being able to predict if a small molecule could bind to one or more protein targets or if a protein could bind to some given ligands is very useful for drug discovery endeavors, anticipation of binding to off- and anti-targets [1]. To date, several studies explore such questions from chemogenomic approach to reverse docking methods [2-5]. Most of these studies have been performed either from the viewpoint of the ligands or of the targets. However it seems valuable to use information from both ligands and binding pockets [6-9]. Our objective is to propose a fine characterization of the protein-ligand interaction based on both ligand and pocket description in a perspective to predict the other partner by knowing one of the interaction partners.

Three types of properties or 3D molecular descriptors are classically used in structural bioinformatics and chemometrics to describe the ligand space or the pocket space : (i) geometric descriptors such as volume, shape and narrowness, (ii) physico-chemical descriptors such as polarity, hydrophobicity or (iii) pharmacophore approaches. Pharmacophore represents the spatial arrangement of features that enables a ligand to interact with a protein in a specific binding mode. With a pharmacophore approach, atoms are described by the properties they have acquired by their spatial arrangement in the molecule. Various ligand-based and structure-based methods have been developed to improve pharmacophore modeling and have been successfully applied in virtual screening, de novo design and lead optimization [10]. Recently, some pharmacophore models have been derived from protein-ligand 3D complex structures in a perspective of linking ligands to putative targets [11]. In this study, we will explore the complementarity of geometric and physico-chemical descriptors with pharmacophore approach.

Our dataset is composed of 483 structures of protein-ligand complexes with a resolution better than 2.5 Å. In a first step, pocket and ligands pair were characterized by combining physico-chemical descriptors and geometric descriptors, optimized on both pocket and ligand spaces. Then, a Principal Component Analysis (PCA) was performed on a combination of 24 descriptors and enables an in-depth study of the properties shared by similar pairs and providing a suitable representation of the pocket-ligand pair space [12]. This PCA analysis shows that neither pocket descriptors nor ligand descriptors gather the whole information. This validates the fact that it is more informative to take into account pocket space and ligand space together than independently whenever possible. To map and group pocket-ligand pairs with similar properties, we built a hierarchical clustering tree on the space defined by the PCA obtained on our set of 483 complexes [12]. This analysis provides a detailed classification of eight particular types of pocket-ligand pairs exibiting some specific properties and highlights strong correspondences between the two interacting partners. For example, we observe that pockets with high roughness values tend to bind ligands with weak polar surface area, hydrogen bond acceptor and donor counts values, and therefore a lower polarity.

In a second step, the aim is to check if the eight pocket-ligand pairs clusters can be complementary characterized by interaction properties given by pharmacophore approach. For this purpose, we will compute pharmacophores based either on pocket and/or ligand. The input and complementary of these pharmacophores will be tested for similar complexes based on specific families of proteins (eg, protein kinase). Then, they will be extended to all of the eight pocket-ligand pairs clusters.

The information to characterize the different partners could be used in a perspective of pocket or ligand profiling, i.e. to predict some ligand properties critical for binding to a given pocket, and conversely, some key pocket properties for ligand binding.

References

[1] S. Perot, O. Sperandio, M. Miteva, A.C. Camproux, B. Villoutreix, Druggable pockets and binding site centric chemical space : a paradigm shift in drug discovery. *Drug Discovery Today.* 15:656-67, 2010.

[2] J. Mestres, Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Current Opinion in Drug Discovery & Development.* 7: 304-313, 2004.

[3] M.J. Keiser, B.L. Roth, B.N. Armbruster, P. Ernsberger, J.J. Irwin, et al., Relating protein pharmacology by ligand chemistry. *Nature Biotechnology.* 25: 197-206, 2007.

[4] Y.Z. Chen, C.Y. Ung, Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. *Journal of Molecular Graphics & Modelling.* 20: 199-218, 2001.

[5] G. Wolber, T. Seidel, F. Bendix, T. Langer, Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today.* 13: 23-29, 2008

[6] Y. Yamanishi, E. Pauwels, H. Saigo, V. Stoven, Extracting sets of chemical substructures and protein domains governing Drug-Target interactions. *Journal of Chemical Information and Modeling.* 51: 1183-1194, 2011

[7] M. Lapinsh, P. Prusis, S. Uhln, J.E.S Wikberg, Improved approach for proteochemometrics modeling: application to organic compound-amine G protein-coupled receptor interactions. *Bioinformatics.* 21: 4289-4296, 2005.

[8] N. Weill, D. Rognan, Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *Journal of Chemical Information and Modeling.* 49: 1049-1062, 2009.

[9] J. Meslamani, D. Rognan, Enhancing the accuracy of chemogenomic models with a three-dimensional binding site kernel. *Journal of Chemical Information and Modeling.* 51: 1593-1603, 2011.

[10] S.Y. Yang. Pharmacophore modeling and applications in drug discovery : challenges and recent advances. *Drug Discovery Today.* 15:444-50, 2010.

[11] J. Meslamani, J. Li, J. Sutter, A. Stevens, H.O. Bertrand, D. Rognan, Protein-ligand-based pharmacophores : generation and utility assessment in computational ligand profiling. *Journal of Chemical Information and Modeling.* 52: 943-955, 2012.

[12] S. Perot, L. Regad, C. Reynes, O. Sperandio, M. Miteva, B. Villoutreix, A.C. Camproux, Insights into an original pocket-ligand pair classification: a promising tool for ligand profile prediction. *Plos One. in press*.

Population genomics of a protoploid yeast species: *Saccharomyces kluyveri*

Python tools development for Next Generation Sequencing analyses

Sophie SIGUENZA[1], Anne FRIEDRICH[1], Cyrielle REISSER[1], Paul JUNG[1] and Joseph SCHACHERER[1]

[1] DEPARTMENT OF GENETICS, GENOMICS AND MICROBIOLOGY, UMR7156, University of Strasbourg / CNRS, 28 rue Goethe, 67083, Strasbourg, Cedex, France
{siguenza, anne.friedrich ,creisser, pauljung, schacherer}@unistra.fr

Keywords: Population genomics, Next Generation Sequencing, Python

Population genomics of a protoploid yeast species: *Saccharomyces kluyveri*

Python Tools development for Next Generation Sequencing analyses

Keywords: Population genomics, Next Generation Sequencing, Python.

1 Introduction

The genetic variation that occurs naturally in a population represents a unique resource for studying the genetic basis of phenotypic diversity between individuals. In this context, yeasts are of particular interest as phenotypic diversity among isolates is significant at different levels. To date, yeast population genomics focused on two closely related species: *S. cerevisiae* and *S. paradoxus* [1][2]. Interestingly, the *Saccharomyces* genus underwent whole-genome duplication, which had an important impact on the evolution of the genomes. In this context, we launched a comprehensive survey of genomic and phenotypic variations among isolates within a protoploid yeast species, which did not undergo whole genome duplication: *Saccharomyces kluyveri*.

2 Next Generation Sequencing Analyses

2.1 Reference sequence, data set and tools for analyses

The genome of a reference *S. kluyveri* strain (CBS 3082) has already been completely sequenced and annotated, providing access to a high quality reference genome. To explore the genetic diversity and to compare the patterns of DNA sequence variation, we sequenced the genome of a large collection of *S. kluyveri* strains isolated from various geographical locations (North America, Asia and Europe) and ecological niches (*Drosophila*, tree exudates, soil).

To study variations within this specie, we sequenced the whole genomes of 29 strains. The DNA of each isolate has been sequenced with Next Generation Sequencing, using Pair end Illumina HiSeq 2000. Statistical analysis and reads quality control have been processed with FastQC [3]. Reads mapping have been performed against CBS 3082 reference sequence with BWA [4] while *de novo* assemblies have been constructed with SOAPdenovo [5], in order to perform structural variations studies. Finally, SNPs and small-indels calling have been realized thanks to SAMtools [6]

We found that these genomes are highly polymorphic. The phylogeny and population structure of *S. kluyveri* provide clear evidence for well-defined geographically isolated lineages. In addition, the systematic screen of the growth rate variation in 60 conditions shed on light a broad phenotypic diversity.

Our analyses will allow us to have a deeper insight into the genetic basis of phenotypic diversity within a protoploid yeast species.

2.2 Python development Tools

Beyond the fact that Python is historically the team's scripting language, it reveals many advantages as well as for quick analyses as for tools development in this project context. Python is extensible, flexible and allows sequential and object programming. In addition, many libraries have already been developed for scientific usages (ie: Biopython [7]), easing the development of our own libraries. Many aspects of python programming have been used to answer our needs. The priority has been given to pipe the NGS data analyses while developing a POO Mutation Discovery library in background. We also used data driven analysis to automatically launch complex processes guided by metadata generation. Finally, reproducibility and confidence in treatment encouraged us to develop reusable and "readable" scripts or POO modules.

Acknowledgements

This work is supported by AGENCE NATIONAL DE LA RECHERCHE, CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE, UNIVERSITÉ DE STRASBOURG, REGION ALSACE (FRANCE) and by NATIONAL INSTITUTES OF HEALTH (UNITED STATES).

References

[1] J. Schacherer, J.A. Shapiro, D.M. Ruderfer and L. Kruglyak, Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. Nature, 458(7236):342-345, 2009.

[2] G. Liti G, D.M. Carter, A.M. Moses, J. Warringer, L. Parts, S.A. James, R.P. Davey, I.N. Roberts, A. Burt, V. Koufopanou , I.J. Tsai, C.M. Bergman, D. Bensasson, M.J. O'Kelly, A. van Oudenaarden, D.B. Barton, E. Bailes, A.N. Nguyen, M. Jones, M.A. Quail, I. Goodhead, S. Sims, F. Smith, A. Blomberg, R. Durbin and E.J. Louis. Population genomics of domestic and wild yeasts. Nature, 458(7236):337-341, 2009

[3] Andrew S. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[4] Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-1760. [PMID: 19451168]

[5] Luo et al.: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 2012 1:18.

[6] Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

[7] Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, and de Hoon MJ. *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics 2009 Jun 1; 25(11) 1422-3. doi:10.1093/bioinformatics/btp163 [PMID:19304878]

A new method for spectral characterization of protein families from sequence information using Fourier Transform

Magali BERLAND[1], Bernard OFFMANN[2,3], Isabelle ANDRÉ[4], Magali REMAUD-SIMÉON[4] and Philippe CHARTON[1]

[1] DSIMB, UMR S-665, INSERM, 15 av. René Cassin, Université de La Réunion, 97715 St Denis, Cedex 09, France
[2] UFIP, FRE 3478 CNRS, 2 chemin de la Houssinière, Université de Nantes, 44322 Nantes, Cedex 3, France
[3] Peaccel SAS, Saint Denis, La Réunion
[4] INSA, UPS, INP; LISBP, UMR5504 CNRS, UMR792 INRA, 35 Avenue de Rangueil, Université de Toulouse, 31077, Toulouse, France
{magali.berland, bernard.offmann}@univ-nantes.fr

Keywords protein sequence analysis, rrm, signal processing, consensus spectrum, bootstrap.

1 Introduction

The Resonant Recognition Model (RRM) is a physico-mathematical model [1] that tentatively links the long-range electron charge transfer along the protein backbone to the biological properties of proteins. This process attributes to each amino acid a numerical electron-ion interaction potential (EIIP) calculated on the base of their valency electrons. The RRM provides a framework to identify, from sequence information, original features extracted through their numerical encoding and subsequent treatment by Fast Fourier transform (Fig. 1a). Proteins that are evolutionarily related are supposed to share common spectral features, which are fingerprints of the biological function [2], [5], [3]. According to the current RRM procedure [4], the consensus spectrum is obtained by simply multiplying the individual spectrum of the family members, but it is highly sensitive to the representativity of the sequences available for the family. Besides, the prominent peaks, corresponding to the "characteristic" frequencies, are identified in the RRM method by applying an arbitrary signal-to-noise (s/n) cut-off value of 20. But this method suffers two main flaws: the s/n values are highly dependent on the number of proteins in the family, so (i) the chosen cut-off may sometimes never be reached, and (ii) it is difficult to compare consensus spectra from functional families of different sizes. Furthermore, the RRM method does not propose to evaluate the statistical significance of those prominent peaks. This work aims to address these issues.

1.1 Method

We provide a new method for calculating and characterizing the consensus spectra in the RRM method. Our methodology proceeds by calculating the geometric mean of the energy spectra of sequences from a functionally characterized group of sequences, and a logarithmic transformation to ensure the conservation of all the relevant information, in the case of the families containing a large number of proteins. The amplitude of the recurrent peaks hence the s/n ratio are, in our method, independent of the number of sequences in the initial dataset because we take the nth root of the cross spectra. Henceforth, to evaluate the significance of a peak, we propose a s/n cut-off value that is based on the standard deviation of the s/n values to determine characteristic peaks of the consensus spectrum. The statistical significance of each of these peaks was evaluated using a bootstrap strategy which consists in shuffling 1000 times the amino acid order for each sequence of the family.

1.2 Result

We illustrate this new canvas on the analysis of a highly divergent Classic odorant binding proteins from three species of mosquitoes, and of a more conserved sialidase enzyme family from trypanosomes. As shown on Fig. 1b the s/n ratio of prominent peaks in the RRM method is clearly dependent on the number of sequences whereas, when the geometric mean is used, the s/n ratio is not influenced by the number of sequences.

We calculated the percentage of random peak occurrence from the shuffled sequences for a range of thresholds based on the standard deviation of the signal-to-noise values (Fig. 2a and Fig. 2b). When this percentage drops below 1%, the corresponding threshold is chosen for determining the prominent peaks and characteristic frequencies. The bootstrapping strategy used provided an assessment of the statistical significance of all these peaks (Fig. 2c).

The proposed methodology provides a rational framework for a robust spectral characterization of protein sequences using Fourier transform. The application of such a method for automated classification and for protein design promises exciting breakthroughs.

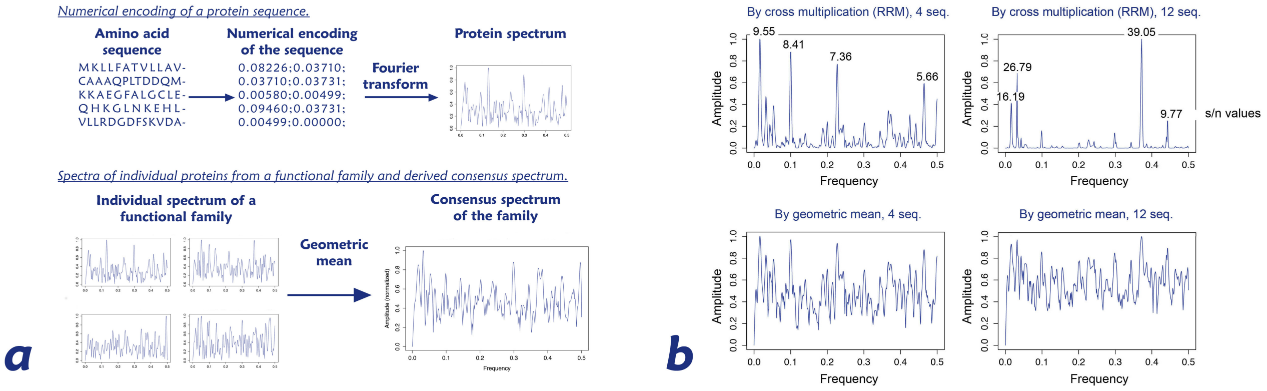


Figure 1. (a) Principle of RRM method: Numerical encoding of a protein sequence and consensus spectrum calculation from individual spectra from a functional protein family. (b) Consensus spectra representations: Spectra were calculated by cross multiplication and geometric mean showing the effect of the number of sequences.

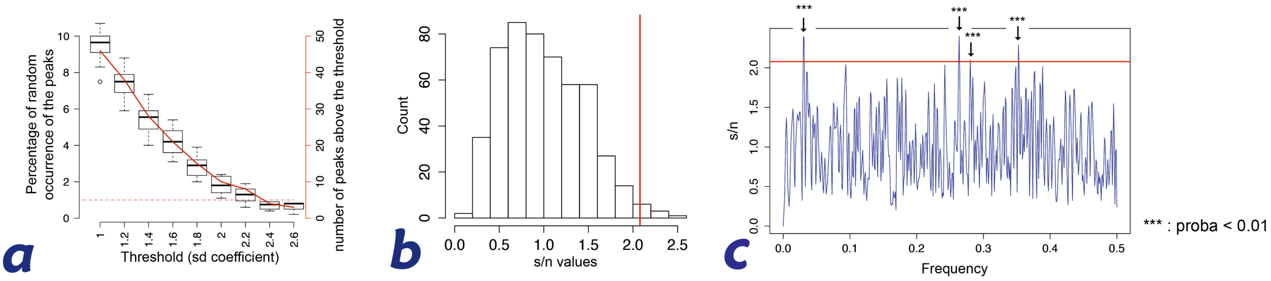


Figure 2. Characterization of the prominent peaks in a consensus spectrum (catalytic domain of a sialidase). (a) Percentage of random peak occurrence as a function of the standard deviation based cut-off value. (b) Distribution of s/n values and the used cut-off to select prominent peaks. (c) Consensus spectrum showing s/n values of peaks above the cut off and their statistical significance as assessed by the bootstrap strategy.

References

[1] Irena Cosic. *The resonant recognition model of macromolecular bioactivity*. theory and applications. Birkhauser, 1997.

[2] Irena Cosic and Elena Pirogova. Bioactive peptide design using the Resonant Recognition Model. *Nonlinear Biomedical Physics*, 1(1):7, 2007.

[3] Taghrid S Istivan, Elena Pirogova, Emily Gan, Nahlah M Almansour, Peter J Coloe, and Irena Cosic. Biological Effects of a De Novo Designed Myxoma Virus Peptide Analogue: Evaluation of Cytotoxicity on Tumor Cells. *PLoS ONE*, 6(9):e24809, September 2011.

[4] Norbert Nwankwo. Digital Signal Processing Techniques:Calculating Biological Functionalities. *Journal of Proteomics & Bioinformatics*, 04(12), 2012.

[5] E Pirogova, V Vojisavljevic, J L Hernández Cáceres, and I Cosic. Ataxin active site determination using spectral distribution of electron ion interaction potentials of amino acids. *Medical & Biological Engineering & Computing*, 48(4):303–309, February 2010.

Druggability pocket prediction from apo and holo proteins

Alexandre BORREL[1,2,3], Leslie REGAD[1,2], Henri XHAARD[3], Michel PETITJEAN[1,2] and Anne-Claude CAMPROUX[1,2]

[1] INSERM, U973, F-75205 Paris, France

[2] Univ Paris Diderot, Sorbonne Paris Cité, UMRS 973, MT*i*, F-75205 Paris, France
{alexandre.borrel, leslie.regad, michel.petitjean ,anne-claude.camproux}@univ-paris-diderot.fr

[3] Centre for Drug Research, Faculty of Pharmacy, FI-00014 University of Helsinki, Finland
{alexandre.borrel, henri.xhaard}@helsinki.fi

Keywords druggability, predictive model, drug discovery.

Therapeutical molecules bind to preferred sites of action, which are in the majority of cases pockets located within proteins or at their surface. Therefore, estimation and characterization of pockets is a major issue in drug target discovery. Among the molecules, “drug-like molecules” [1] are small molecules with particular properties as of small size, able to cross the digestive tract. They represent a large number of drugs present in current pharmacopoeia such as for examples, toxol or morphin. Pocket “druggability”, the ability of a pocket to bind “drug-like” molecule, is essential for drug discovery studies [2] especially for discovering new targets.

Currently, identifying druggable pockets is possible by different statistical models of prediction [3, 4]. These methods differ by methods used to estimate pockets, by descriptors used to characterize pockets and the statistical methods used. Moreover, the quality of these approaches is limited by the limited amount of data available, and most of them allow the prediction of the pocket druggability if the structure of the target is complexed to one ligand (holo forms). However in order to discover novel targets, it is important to be able to predict the “druggability” of a pocket in its apo form, i.e. not yet bound and deformed by the interaction with a ligand. Here, we propose two models to predict pocket druggability: one allows the druggability prediction of pocket in its holo form and the second model allows the druggability prediction of pocket from its apo form. To develop these two models, we started from a set of 113 complexes protein ligands [3], with 71 druggable proteins and 41 less druggable proteins. From this set, we used two approaches to estimate pockets by taking or not the ligand information. The first allows the estimation of pockets on holo form of the protein and defines pockets as protein atoms less than 4 Å away from the ligand. The second method estimates pockets without information about ligands and based on the geometric algorithm Fpocket [5]. This second approach allows the estimation of pocket on both holo and apo forms of proteins.

Pockets estimated using two approaches, were then characterized using a set of 57 descriptors. This descriptor set, named pocket profile, allows a characterization of the geometry and the physicochemical properties of pockets [6, 7, 8].

We then built statistical models based on a linear discriminant analysis to predict the pocket druggability from the two datasets of pockets based on their pocket profile. The first model trained on descriptors computed on pockets estimated using the ligand information allowed us to develop a new method of druggability prediction of pockets in holo form. The second model based on descriptors computed for pockets estimated using Fpocket allowed us to predict druggability of pockets in their holo and apo forms. The construction of these models consisted in the selection of the models with the best accuracy and containing as few descriptors as possible. Our two models present a very good accuracy (close to 80%). The model trained on pockets estimated using ligand information contains 4 descriptors: 2 physicochemical descriptors reflecting the pocket hydrophobicity and the number of aromatic residues and 2 geometric descriptors. The performance of this model is similar to other known models and the selected descriptors are in agreement with these models [7, 9, 11]. Our second model, trained on the pockets estimated without ligand information, is more effective than other studies [5]. It is based on 3 physicochemical descriptors: 2 are common with our first model. Thus, we can conclude that our two models are dependent on the pocket estimation, showing the interest of the two models. Since the estimation with Fpocket is not based on the ligand information, the model trained on pocket estimated by this approach can predict the druggability of pockets extracted from proteins in apo form. While our dataset used for the model training corresponds to pockets extracted from holo form, however, the binding of a ligand can cause a deformation of the protein. Thus, in order to analyses of the deformation effect on our model, we tested the second model on a set of apo pockets. On this new dataset, the model based on Fpocket estimator gives a good accuracy (close to 90%). This suggests that this model is able to capture druggability characteristics of pockets from its apo form if global properties of a pocket are conserved between apo and holo forms. The perspective to pocket druggability prediction is when only the structure of the target in the apo form is available, which in new target discovery.

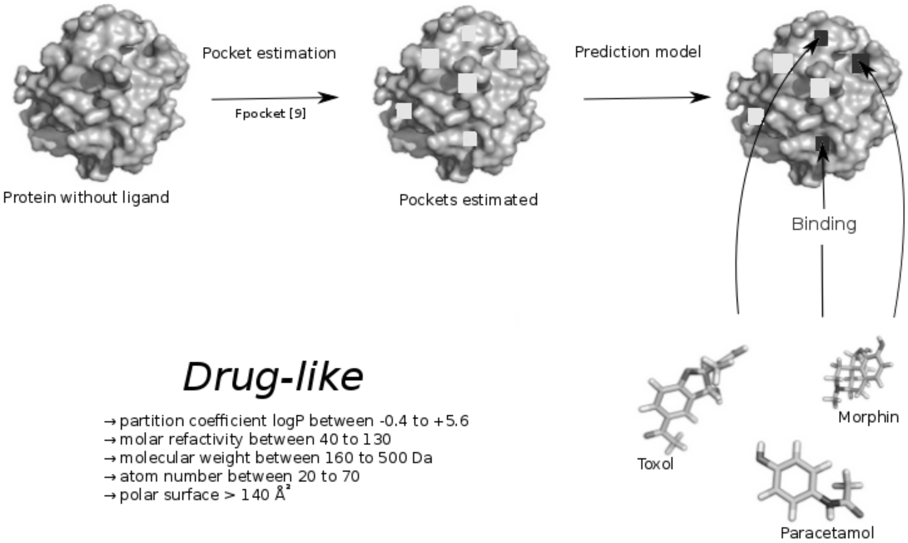

Pocket estimation
Fpocket [9]
Prediction model
Protein without ligand
Pockets estimated
Binding
Drug-like
→ partition coefficient logP between -0.4 to +5.6
→ molar refactivity between 40 to 130
→ molecular weight between 160 to 500 Da
→ atom number between 20 to 70
→ polar surface > 140 Å²
Toxol
Morphin
Paracetamol

Figure 1: The aim of this study is to predict if a protein pocket is able to bind a drug-like molecule.

References

[1] Lipinski, C., Lombardo, F., Dominy, B. W., & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews, 46(1-3), 3–26, 2001

[2] Hann, M., & Keserü, G. Finding the sweet spot: the role of nature and nurture in medicinal chemistry. *Nature Reviews Drug Discovery, 11, 355-365*, 2012

[3] Perola, E., Herman, L., & Weiss, J. Development of a Rule-Based Method for the Assessment of Protein Druggability. *Journal of chemical information and modeling, 52(4), 1027–1038,* 2012

[4] Milletti, F., & Vulpetti, A. Predicting poly-pharmacology by binding site similarity: from kinases to the protein universe. *Journal of chemical information and modeling,* 50(8), 1418–31, 2010

[5] Eyrisch, S., & Helms, V. Transient pockets on protein surfaces involved in protein-protein interaction. *Journal of medicinal chemistry,* 50(15), 3457–64, 2007

[6] Kyte, J., & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology,* 157(1), 105–32. 1982

[7] Nisius, B., Sha, F., & Gohlke, H. Structure-based computational analysis of protein binding sites for function and druggability prediction. *Journal of biotechnology,* 159(3), 123–134, 2011

[8] Petitjean, M. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *Journal of Chemical Information and Computer Sciences* 32(4), 331–337, 1992

[9] Le Guilloux, V., Schmidtke, P., & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics,* 10, 168, 2009

[10] Krasowski, A., Muthas, D., Sarkar, A., Schmitt, S., Brenk, R. DrugPred: a structure-based approach to predict protein druggability developed using an extensive non redundant data set. *Journal of chemical information and modeling,* 51(11), 2829–42, 2011

[11] Desaphy, J., Azdimousa, K., Kellenberger, E., & Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *Journal of chemical information and modeling,* 52(8), 2287–99, 2012

Moonlighting proteins: Not as uncommon as assumed

Charles E. CHAPPLE[1], Benoit ROBISSON[1] and Christine BRUN[1]

TAGC Inserm U928, Université de la Méditerranée, Parc Technologique de Luminy,
Case 928, 13288, Marseille, Cedex 9, France
{cchapple},{robisson},{brun},@tagc.univ-mrs.fr

Keywords PPI network, moonlighting, GeneOntology, protein function, graph partitioning.

1 Introduction

Piatigorsky *et al* [1] identified the first moonlighting protein in 1988 when they observed that the lens structural protein delta-crystallin was also an enzyme. They called this phenomenon "gene sharing" but "protein moonlighting" has since been adopted instead. We now know that multifunctional proteins are far more common than initially assumed, and therefore define moonlighting proteins as the subset of multifunctional proteins whose multiple functions are autonomous, unrelated, and not necessarily partitioned into different protein domains. Human aconitase, for example, is an enzyme of the TCA cycle but is also a translational regulator [2].

The aim of this project is the large scale prediction of moonlighting proteins from PPI networks. Our first step is covering our PPI network with a system of overlapping clusters. Proteins in the same clusters tend to be involved in similar biological processes and those belonging to **multiple clusters** are likely to have **multiple functions**. Therefore, proteins found at the intersection of or otherwise linking clusters involved in different processes are good moonlighting candidates.

2 Methods

2.1 Networks

We have built a non-redundant, connected, high quality human PPI network of 12865 nodes and 74388 edges by combining data from multiple online databases. We only kept interactions identified using experimental methods that find direct, binary interactions and have removed redundancy and self interactions.

2.2 Cluster Identification and annotation

We cover our PPI network with a system of overlapping clusters using OCG[3], a community detection algorithm that allows nodes to belong to multiple clusters. Each cluster is then annotated according to the Gene Ontology [4] (GO) annotations of its constituent proteins. A class will be annotated to a specific GO term *iff* $\geq 50\%$ of annotated proteins in that class share that GO term. A class is considered for annotation *iff* at least two thirds of its constituent proteins have GO annotations. All member proteins will inherit the annotation(s) of the class.

2.3 GO probabilities

We have developed two GO term similarity metrics, the probabilities of annotation and interaction. For the first metric, the annotation probabilities, we consider two GO terms to be dissimilar if fewer gene products are directly annotated to both terms than would be expected by chance. For the second metric, the interaction probabilities, we consider two GO terms X and Y to be dissimilar if there are fewer interactions between proteins annotated to X and proteins annotated to Y in a given network than expected by chance.

2.4 MoonGO

MoonGO, our moonlighting protein prediction tool, uses the annotated classes and GO term association probabilities to search the network for proteins found connecting classes annotated to dissimilar GOs. In "intersection" mode, it will search for proteins that are members of both classes and in "bridge" mode for proteins that, while belonging to neither class, have interacting partners in both and are the only link between these two classes. If the candidate is not the only node connecting the classes, it will be discarded.

3 Results

We have found 412 Intersection candidates and 1055 Bridge candidates with 74 proteins identified as candidates by both methods. We then analyzed various features of our candidates to better characterize them (see Table 1).

	Degree	Betweeness	Shortest Paths	Clusters	Annotations (P)	Domains	Disorder	Expression	Isoforms	Length
Intersection	⇑	⇑	⇓	⇑	⇑	⇑		⇑	⇑	⇑
Bridge	⇑	⇑	⇓	⇓	⇓		⇑	⇑	⇑	⇑

Table 1. Comparison of the mean values of different features of the candidate and multi-clustered non-candidate nodes (Multi NC). ⇑ indicates that the candidates had a significantly higher mean value than the Multi NCs and ⇓ that the candidates had a significantly lower mean value than the Multi NCs. ↓ and ↑ indicate non-significant differences in mean values. Blanks indicate no significant differences. The means were compared using the Wilcoxon signed-rank test and the significance threshold is ≤ 0.05. The characteristics examined are the number of interactors (degree), network centrality (betweeness), node interconnectivity (shortest paths), the number of partition clusters, GO biological process annotations, PfamA+B domains, protein disorder, RNA expression, number of isoforms and protein length.

4 Discussion

Our candidates form a novel sub-type of multifunctional proteins that differ significantly from other multifunctional proteins. At first glance they behave like hubs only more so, a trend that is also observed in yeast, indeed 361 of our 412 (88%) candidates are hubs though only 15% of hubs are candidates. However, they are significantly less disordered than the hub average and they are also less ubiquitously expressed at the protein level, suggesting a possible functional constraint that separates our candidates from the other hubs. Surprisingly, we have identified 10% of the network as moonlighting proteins indicating that such extreme multifunctionality may be much more common than previously assumed.

Acknowledgements

This work was funded by an ANR grant to C.B., B.R. is a PhD fellow of the AXA Research Fund and CEC is a postdoctoral fellow of the Fondation pour la Recherche Médicale.

References

[1] J. Piatigorsky, W. E. O'Brien, B. L. Norman, K. Kalumuck, G. J. Wistow, T. Borras, J. M. Nickerson, and E. F. Wawrousek. Gene sharing by delta-crystallin and argininosuccinate lyase. *Proc Natl Acad Sci U S A*, 85(10):3479–3483, May 1988.

[2] Karl Volz. The functional duality of iron regulatory protein 1. *Curr Opin Struct Biol*, 18(1):106–111, Feb 2008.

[3] Emmanuelle Becker, Benoît Robisson, Charles E Chapple, Alain Guénoche, and Christine Brun. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, 28(1):84–90, Jan 2012.

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

Diversity of CRISPR systems in the euryarchaeal Pyrococcales

Cédric NORAIS[1], Annick MOISAN[2], Christine GASPIN[2] and Béatrice CLOUET D'ORVAL[3]

[1] LABORATOIRE DE BIOCHIMIE, UMR7654 CNRS, Ecole Polytechnique, Palaiseau, France
Cedric.Norais@polytechnique.edu

[2] LABORATOIRE DE MATHEMATIQUE ET INFORMATIQUE APPLIQUEES, UR875 INRA, Unité de Biométrie et Intelligence Artificielle, F-31326, Castanet-Tolosan, France.
{Annick.Moisan, Christine.Gaspin}@toulouse.inra.fr

[3]LABORATOIRE DE MICROBIOLOGIE ET GÉNÉTIQUE MOLÉCULAIRE, UMR 5100, CNRS et Université de Toulouse III, 31062 Toulouse, France.
Beatrice.Clouet-dorval@ibcg.biotoul.fr

Abstract *Pyrococcales are a group of hyperthermophilic euryarchaea that are frequently found in deep sea hydrothermal vents. Infectious genetic elements, such as plasmids and viruses, remain a threat even in this remote environment and these microorganisms have developed several ways to fight their genetic invaders. Among these are the recently discovered CRISPR systems. In this analysis, we have combined and condensed available information on CRISPR systems found in the Pyrococcales to fight them.*

Keywords Archaea, CRISPR system, ncRNA.

1 Introduction

Pyrococcales are a group of hyperthermophilic euryarchaea that are frequently found in deep sea hydrothermal vents. Infectious genetic elements, such as plasmids and viruses, remain a threat even in this remote environment and these microorganisms have developed several ways to fight their genetic invaders. Among these are the recently discovered CRISPR (clustered regularly interspaced short palindromic repeats) systems. CRISPR loci are composed of arrays of direct repeats separated by variable sequences called spacers or guides that generally derived from invader genetic elements. A match between a CRISPR guide and an invading nucleic acid provides immunity to infection. A dozen major groups of *cas* (CRISPR-associated) gene families, located near CRISPR loci are proposed to be involved in the three major phases of CRISPR systems: adaptation, expression, interference. In this analysis, we have combined and condensed available information on CRISPR/Cas systems found in the Pyrococcales. The organization of CRISPR/*cas* systems will be presented according to the nomenclature proposed by Makarova et al [2] with emphasis on the genomic arrangements of CRISPR arrays and *cas* modules within pyrococcal genomes.

2 Results and discussion

Three different effector *cas* gene types, defined by the presence of specific signature genes, were identified within the six completely sequenced pyrococcal genomes: a subtype I-A, a subtype III-A, and a subtype III-B system. Typically, subtype I-A is ubiquitously represented in Pyroccocales. However, a closer look to CRISPR arrays and core gene sequences suggests that pyrococcal CRISPR systems fall into two categories, associated either with subtype I-A or III-A systems. Each group seems to have its own set of leader sequences and direct repeats, informational module, crRNA processing Cas6, and effector Cas complexes that are separated into multiple modules. Multiple effector clusters are found associated to subtype I-A systems, including the I-A_1 and I-A_2 modules. These gene clusters seem to be two homologous subtype I-A effector modules, characterized by the presence of either *cas8a1/cst1* or *cas8a2*. We suggest that subtype III-B genes might be viewed as a subtype I-A effector module, rather than an autonomous CRISPR/*cas* system. For instance, in the case of pyrococcal genomes, no informational module or CRISPR array are ever found associated to subtype III-B gene clusters, except when they form a subtype I-A/III-B hybrid system where they are associated to the subtype I-A informational module and CRISPR array (as found in *P. furiosus*). As opposed to all other CRISPR systems that target DNA, subtype III-B effector RNP complex is the only one known to target RNA molecules. This subtype III-B module could be viewed as an option, for the subtype I-A CRISPR system, conferring the ability to target RNA in addition to DNA, using the same set of guide sequences. The existence of such an RNA-targeting system is intriguing as no archaeal RNA virus has been identified so far. This raises the question: why has such an RNA targeting module developed? Is there a benefit for a hyperthermophilic archaeal cell to degrade RNA molecules from an invader? There might be some RNA viruses from deep vents that we have not yet identified. This system could alter the expression of infecting genetic elements by silencing their mRNAs while other defence systems (CRISPR or not) rid the cell from the invader or triggers programmed cell death, thus preventing the infectious genetic element from spreading across the population. The subtype III-B system could also have functions other than anti-viral defense, such as gene regulation, although guides targeting cellular genes are the exception rather than the rule. Although no experimental evidence exists so far, subtype I-A and subtype III-B are hypothesized to use the same set of crRNAs (from group 1 CRISPR arrays), whereas Type III-A complexes seem to be independent and to use their own set of crRNAs (from group 2 CRISPR arrays, which are absent from *P. furiosus* and *P.* species NA2 genomes). Group 1 and group 2 crRNAs contain different 5' tags that hypothetically allow each effector complex to select its dedicated crRNA after their processing by Group 1 or 2 specific Cas6 endoribonucleases. The *P. abyssi* genome differs from the other Pyrococcales. It is intriguing because it does not present a single full CRISPR system. The *P. abyssi* genome contains the I-A_2 effector module but lacks the associated informational module and more specifically *cas1* and *cas2* genes, which theoretically impairs this species from acquiring new guide sequences. *P. abyssi* carries a group 2 CRISPR array, indicating it might have once possessed a functional subtype III-A system, which could have been lost afterward. Thus the intriguing *P. abyssi* CRISPR profile is most probably the result of gene loss. This organism possibly once had a more elaborated CRISPR system organization and might have lost several of its CRISPR components. This raises the question of whether *P. abyssi* has CRISPR activity. Experimental evidences show that two of the four CRISPR arrays are expressed and their transcripts processed into individual crRNAs. These two arrays are of the subtype I-A group for which *cas* genes (I-A_2 module) are still present in the strain. Thus it is probable that this residual CRISPR system can still target invaders bearing some complementarity to the set of guides from expressed arrays but cannot develop immunity towards novel genetic invaders, since *cas1* and *cas2* appear to miss.

3 Conclusion

The origins of pyrococcal CRISPR guide sequences remain to be determined. The lack of matching hits certainly results from the scarcity of available thermococcal plasmid and virus sequences. The ever increasing number of available sequences will at some point allow the discovery of other matches between Thermococcal CRISPR guides and genetic invaders, providing additional evidence that CRISPR systems participate in the defence against plasmids and viruses.

The diversity of CRISPR system situations in the Pyrococcales illustrates the modularity and instability of CRISPR systems. It would be quite challenging to decipher the ancestral CRISPR organization in the Pyrococcales. One possibility is that the common ancestor to all Pyrococcales had all three CRISPR subtype systems, which would have then been differentially lost over time, depending on the species (e.g. *P. abyssi* or *P.* species NA2). The opposite possibility is that the common ancestor had no CRISPR at all and species have been differentially "infected" (inseminated?) by various CRISPR systems that could have been brought to them by mobile genetic elements, such as plasmids, which have been reported to also carry CRISPR systems.

Thus CRISPR systems in the Pyrococcales are diverse and dynamic. There are a lot of things that we still do not understand on the pyrococcal CRISPR/*cas* systems organization. But knowledge in the CRISPR field is added as quickly as guide sequences in an active CRISPR array.

References

[1] Norais C, Moisan A, Gaspin C, Clouet-d'Orval B. Diversity of CRISPR systems in the euryarchaeal Pyrococcales. RNA Biol. 2013 Feb 19;10(5).

[2] Makarova KS, Haft DH, Barrangou R, Brouns S, Charpentier E, Horvath P, Moineau S, Mojica FJ, Yakunin AF, van de Oost J, et al.: **Evolution and classification of the CRISPR/Cas systems**. *Nat Rev Microbiol* 2011; 9:467-77.

Bioinformatics tools for large protein complex analysis.

Description of cytokine signaling complex assembly and mechanism from human helper T-lymphocytes.

Anne MENARD[1], Thierry ROSE[1] and Pascal BOCHET[1; 2]

[1] Institut Pasteur, Unité d' immunogénétique Cellulaire, 25 rue du Dr Roux, 75724, Paris, Cedex 15, France
menard.annec@gmail.com,{thierry.rose,pascal.bochet}@pasteur.fr

[2] UMR 3525 CNRS, 25 rue du Dr Roux, 75724, Paris, Cedex 15, France
pascal.bochet@pasteur.fr

Abstract *This paper presents bioinformatics tools developed for MS-data analysis of large protein complexes involved in interleukin-7 signaling regulating CD4 T cell. Tools use data analysis already existing programs and add the functionality.*

Keywords signalization, MS data mining, mass spectrometry, protein-protein interaction network

Interleukin-7 (IL7) is a cytokine regulating differentiation, development, survival and proliferation of helper T-lymphocytes (CD4). These CD4 T-cells supervise the immune survey and response. A regulating loop links IL7 production by non-immune cells and the CD4 T-cell count. This loop is broken in HIV-infected patients: CD4 T-cells are the main target of HIV, their response to IL7 is altered, CD4 T-cell count drops and causes immunodeficiency syndrome.

Our team has shown by super resolution microscopy that IL7 induces the compartmentalization of its receptor in a membrane microdomain (MMD) then the anchoring of its cytoplasmic domains to the actin meshwork [2]. Immunoprecipitation of detergent-solubilized receptors pulled out more than one hundred proteins, specifically recruited after IL7 binding [1].

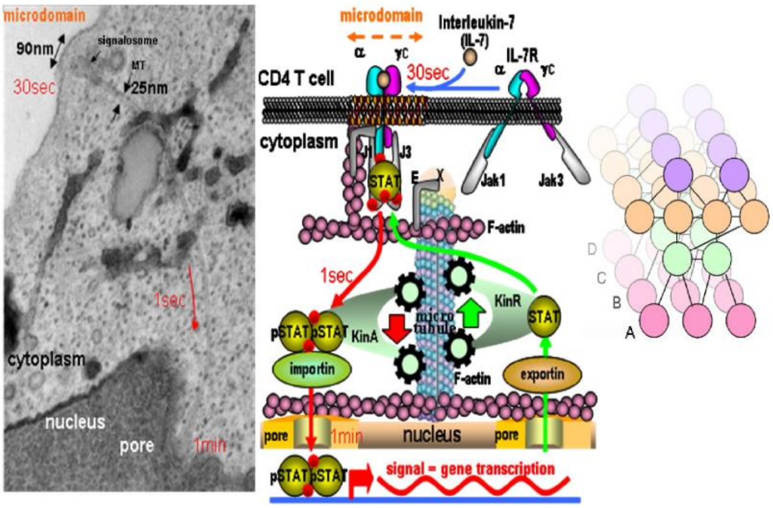


Figure 1. Transmission electronic microscopy image of the CD4 T-cells (left, uranyl and osmium staining) , schematic representation of the signalosome induced by IL-7 (middle) and its representation as Cytoscape layers of protein complex (right). The proteins bound to the IL7-receptor chain alpha forming the signalosome are identified in this study (2013).

We analyzed the protein content of IL7 signaling complexes immunoprecipitated with the IL7-receptor chain alpha (Fig. 1 on the preceding page shows this complex) by LC-MS/MS using an Orbitrap (Proteomics Core Facility, Institut Pasteur). The composition of IL7 signaling complexes was studied at different times from 0 to 15 minutes after receptor activation in the presence of drugs inhibiting MMD formation, actin filament assembly or microtubule polymerization.

We wrote programs building protein lists from MS data analyzed with X! Tandem, Mascot, OMSSA and MS-GF. These programs identify peptides corresponding to digest fragments of the proteins from MS peaks. The lists of proteins and peptides, with details of their identifications, are stored in a database managed with PostgresSQL. Interactive comparisons and boolean operations processed on protein lists using statistics are used to draw hypothesis of assembly mechanism. The comparisons and the visualization are performed on a website in protein lists. Outputs are formatted to be displayed with Cytoscape (www.cytoscape.org). Cystoscape allows the visualization of the co-existence of proteins in immunoprecipitated samples and their Gene Ontology grouping suggesting functional interaction.

Acknowledgements

This work was supported by the “Programme Transversal de Recherche” number 424 from Pasteur Institute.

References

[1] T. Rose, A. Pillet, V. Lavergne, B. Tamarit, P. Lenormand, J.C. Rousselle, A. Namane and J. Thèze, Interleukin-7 compartmentalizes its receptor signaling complex to initiate CD4 T lymphocyte response. *Biol Chem*, 285:14898-14908, 2010.

[2] B. Tamarit, F. Bugault, A. Pillet, V. Lavergne, P. Bochet, N. Garin, U. Schwarz, J. Thèze and T. Rose, Membrane microdomains and cytoskeleton organization shape and regulate the IL7-receptor signalosome in human CD4 T-cells. *J Biol Chem*, 288:8691-701, 2013.

Modeling gamma-Cytokine Signalisation using Smoldyn

Pascal BOCHET[1,2], Blanche TAMARIT[1,3] and Thierry ROSE[1]

[1] Institut Pasteur, Unité d'Immunogénétique Cellulaire, 25 rue du Dr Roux, 75724, Paris, Cedex 15, France
[2] CNRS, UMR 3525 , 25 rue du Docteur Roux, 75724 Paris cedex 15 France.
[3] Université Pierre et Marie Curie, Cellule Pasteur UPMC, 25 rue du Docteur Roux, 75724 Paris cedex 15 France.
{pascal.bochet,blanche.tamarit,thierry.rose}@pasteur.fr

Abstract *We present a simulation of the main signaling mechanism for Interleukin-7 (IL-7) in human helper lymphocytes (CD4+-T cells). The role of active transport on microtubules is evaluated in comparison with the alternative mechanism of diffusion for the movement of the transcription factor STAT5 from the plasma membrane to the nucleus of the cell.*

Keywords receptor, signaling, modeling, protein complex, cytoskeleton.

Helper lymphocytes (CD4+ T cells) regulate the immune response and are among the targets of HIV-1 virus. Gamma-cytokines (interleukin- (IL-) 2, 4, 7, 9, 15 and 21) control the differentiation, the homeostasis and the activation of T lymphocytes. These transmitters bind at subnanomolar concentrations to their respective receptors and signal through common main signaling pathways as Jak/STAT, Pi3K/AKT, MAPK, but play markedly divergent roles in lymphoid biology *in vivo*.

Interleukin signaling events take place within subdomains of the plasma membrane where signaling complexes associated to the receptor are formed [1]. Within these domains the diffusion of signal molecules is restricted, increasing local concentrations and modulating response times. In the Jak/STAT pathway, nuclear translocation of the signaling intermediate, the transcription factor STAT5, is slowed down and reduced in the absence of microdomain formation [2]. Furthermore active transport along components of the cytoskeleton is involved in the translocation of STAT5. The supression of this active transport also slows down the diffusion to the nucleus and reduces the maximum level that it can reach [2].

Smoldyn [3] is a program which simulates the diffusion and the reactions of individual molecules at the cellular scale. However, so far, Smoldyn has no method for the representation of one dimension structures like microtubules or filaments in simulations.

Here we describe an implementation with Smoldyn of hundreds of microtubules bridging the cytoplasmic and nuclear membranes and a simulation for the transport of several thousands signaling molecules carried by unidirectional molecular motors moving along those microtubules. With this simulation we represent signal transduction in helper lymphocytes. The simulation suggests that the transport of STAT5 assisted by molecular-motors along microtubules is in fact slower than simple Brownian diffusion. As microtubules are needed for STAT5 signalisation to occur [2], this renews the question of the role of microtubules in intracellular signaling in this system.

Acknowledgements

Work in the laboratory is supported by the "Programme Transversal de Recherche" number 424 from the Pasteur Institute.

References

[1] T. Rose, A. Pillet, V. Lavergne, B. Tamarit, P. Lenormand, J.C. Rousselle, A. Namane and J. Thèze, Interleukin-7 compartmentalizes its receptor signaling complex to initiate CD4 T lymphocyte response. *J Biol Chem*, 285:14898-14908, 2010.

[2] B. Tamarit, F. Bugault, A. Pillet, V. Lavergne, P. Bochet, N. Garin, U. Schwarz, J. Thèze and T. Rose. Membrane microdomains and cytoskeleton organization shape and regulate the IL7-receptor signalosome in human CD4 T-cells. *J Biol Chem*, 288:8691-8701, 2013.

[3] S. Andrews, N. Addy, R. Brent and A. Arkin. Detailed Simulations of Cell Biology with Smoldyn 2.1, *PLoS Comput Biol* 6:e1000705, 2010.

Identification of new NRPS domain signatures for the prediction of new active cyclic peptides in fungi

Thibault CARADEC[1], Maude PUPIN[2], Malika SMAÏL-TABBONE[3], Marie-Dominique DEVIGNES[3], Philippe JACQUES[1], Valérie LECLERE[1]

[1] ProBioGEM, EA1026, Université de Lille-Sciences et Technologies, Villeneuve d'Ascq, France.
[2] LIFL, UMR8020, Université de Lille-Sciences et Technologies, INRIA Lille-Nord Europe, Villeneuve d'Ascq, France.
[3] LORIA, UMR7503, INRIA NGE, Université de Lorraine, Vandoeuvre-lès-Nancy, France.

Keywords : Nonribosomal peptide, protein signature

La découverte de nouvelles molécules actives représente un challenge permanent, notamment dans le domaine médical. Les métabolites secondaires, produits par les micro-organismes pour répondre à des besoins caractéristiques en matière d'adaptation ou de défense, représentent une source particulièrement intéressante de molécules d'intérêt. Depuis sa découverte dans les années 70, la voie de synthèse peptidique non ribosomique a démontré son potentiel de production de peptides d'intérêt, tels que l'ACV (précurseur de la pénicilline), l'antibiotique bacitracine ou la cyclosporine, un immunomodulateur largement utilisé pour limiter le rejet chez les greffés. La synthèse non ribosomique se démarque du dogme de la biologie moléculaire décrivant la synthèse de protéines comme issue de la transcription d'une séquence d'ADN en ARNm, suivie de la traduction de celui-ci par les ribosomes. La synthèse non ribosomique est réalisée grâce à des lignes d'assemblage appelées NRPS pour *Non Ribosomal Peptide Synthetases* qui sont des complexes multi-enzymatiques modulaires. Chaque module est responsable de l'intégration d'un acide aminé ou monomère dans le peptide produit, et est lui-même divisé en différents domaines enzymatiques. Quatre domaines essentiels ont été décrits, permettant la synthèse peptidique. Le domaine d'adénylation (A) sélectionne et active spécifiquement un monomère par transformation en amino-acyl adénylé, le domaine de thiolation (T) garantit la fixation du peptide en cours de formation sur la synthétase, le domaine de condensation (C) permet la formation de la liaison peptidique entre 2 monomères et le domaine de thioestérase (Te), qui permet le relargage du peptide néoformé et dans certains cas sa cyclisation. Les modules peuvent également inclure divers domaines facultatifs permettant la modification du monomère au cours de la synthèse. Ainsi des domaines d'oxydation (Ox) ou de méthylation (M) peuvent être présents. Mais le domaine de modification le plus largement rencontré est un domaine d'épimérisation (E) aboutissant à la formation d'un monomère d'isomérie D. La voie de synthèse non ribosomique permet ainsi la production de peptides originaux présentant des particularités structurales (peptides cycliques et / ou branchés, présence de monomères non protéogéniques, isomérie D…).

Dans le cadre de la recherche de nouvelles molécules actives, le screening génomique, utilisant des outils bioinformatiques adaptés représente une approche privilégiée. Cette méthode permet notamment de s'affranchir des limites liées à la production de ces molécules, parfois dépendante des conditions de cultures, de la présence d'un substrat spécifique ou d'un autre organisme. Toutefois afin de prédire la structure et l'activité des peptides non ribosomiques, il est nécessaire de bien comprendre les voies de synthèse, particulièrement identifier les activités enzymatique de chaque domaine des NRPS. En effet, l'identification croissante de nouvelles séquences de NRPS tend à démontrer la diversité des activités enzymatique attribuées à chaque domaine.

Ainsi, des alignements de séquences de domaines C et E ont permis de définir plusieurs classes pour chacun de ces domaines et d'établir des relations phylogéniques entre eux [1]. Parmi les domaines C, il existe en fait des domaines C-starters (présents en début de synthèse), des domaines $^{L}C_{L}$ (permettant la condensation entre deux monomères de configuration L) et des domaines $^{D}C_{L}$ (permettant la condensation entre deux monomères d'isoméries différentes). Chez les bactéries du genre *Pseudomonas*, les NRPS impliquées dans la synthèse de lipopeptides cycliques contiennent des domaines bifonctionnels assurant à la fois l'épimérisation et la condensation. Ils ont été appelés Dual C/E. Plus récemment, un domaine C particulier a été identifié chez les champignons. Ce domaine C, en position terminale serait impliqué dans la cyclisation du peptide produit (Ct), remplaçant le domaine Te plus généralement rencontré [2].

Dans un premier temps, nous avons observé que tous ces domaines pouvaient être divisés en deux sous-domaines. Le premier est commun, il correspond en général aux 300 premiers acides aminés (Up-C sequence), alors que le second, spécifique de l'activité enzymatique est porté par les 150 acides aminés suivants (Down-C séquence). Nous nous sommes ensuite focalisés sur les domaines Ct présents uniquement dans les NRPS de peptides cycliques chez les champignons. La base de données NORINE [3] dédiée aux peptides non ribosomiques, a été interrogée pour identifier les peptides cycliques d'origine fongique. L'organisation modulaire et en domaines des synthétases correspondantes a été définie au moyen de logiciels de prédiction spécifiques, tel NRPS-PKS [4]. Au moyen d'outils bioinformatiques, des signatures spécifiques des 150 derniers acides aminés de ces domaines Ct ont été recherchées afin de mieux comprendre le mécanisme d'action de cette sous famille de domaines, et également afin d'établir à l'avenir un moyen simple de les identifier. Pour cela, les séquences ont été alignées au moyen de logiciels d'alignement tels que MUSCLE [5] et ClustalW [6]. La visualisation de signatures a été réalisée par la production de weblogos [7], et une classification phylogénétique a été réalisée via la création d'arbres phylogénétiques au moyen du logiciel MEGA [8]. Une comparaison de ces domaines Ct avec des domaines Te bactériens, qui semblent posséder des activités similaires de cyclisation des peptides, a également été réalisée. L'identification de nouvelles signatures peut permettre une meilleure compréhension des mécanismes de synthèse conduisant à une prédiction plus précise des structures des peptides produits par de nouvelles NRPS identifiées à partir des données de séquençage.

Références

[1] Rausch, C., Hoof, I., Weber, T., Wohlleben, W., and Huson, D.H. (2007). Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. BMC Evolutionary Biology *7*, 78.

[2] Gao, X., Haynes, S.W., Ames, B.D., Wang, P., Vien, L.P., Walsh, C.T., and Tang, Y. (2012). Cyclization of fungal nonribosomal peptides by a terminal condensation-like domain. Nat. Chem. Biol. *8*, 823–830.

[3] Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., and Kucherov, G. (2008). NORINE: a database of nonribosomal peptides. Nucl. Acids Res. *36*, D326–D331.

[4] Ansari, M.Z., Yadav, G., Gokhale, R.S., and Mohanty, D. (2004). NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res. *32*, W405–413.

[5] Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. *32*, 1792–1797.

[6] Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

[7] Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

[8] Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. *28*, 2731–2739.

The Banana Genome Hub

Gaëtan DROC[1†*], Delphine LARIVIÈRE[1,2†], Valentin GUIGNON[3], Nabila YAHIAOUI[1], Dominique THIS[2], Olivier GARSMEUR[1], Alexis DEREEPER[4], Chantal HAMELIN[1], Xavier ARGOUT[1], Jean-François DUFAYARD[1], Juliette LENGELLE[1‡], Franc-Christophe BAURENS[1], Alberto CENCI[3], Bertrand PITOLLAT[1], Angélique D'HONT[1], Manuel RUIZ[1], Mathieu ROUARD[3] and Stéphanie BOCS[1]

[1] AGAP, UMR1334 Cirad, TA A-108 / 03, Av. Agropolis, 34398, Montpellier, Cedex 5, France
{gaetan.droc, delphine.lariviere, nabila.yahiaoui, olivier.garsmeur, chantal.hamelin, xavier.argout, jean-francois.dufayard, franc-christophe.baurens, bertrand.pitollat, angelique.dhont, manuel.ruiz, stephanie.sidibe-bocs}@cirad.fr
jlengelle@gmail.com

[2] AGAP, UMR1334 Montpellier SupAgro, 2 pl Pierre Viala, 34060, Montpellier, Cedex 2, France
dominique.this@supagro.inra.fr

[3] Commodity systems & genetic resources programme, Bioversity International, Parc Scientifique Agropolis II, 1990 Boulevard de la Lironde, 34397, Montpellier, Cedex 5, France
{v.guignon, a.censi, m.rouard}@cgiar.org

[4] RPB, UMR186 IRD, 911 avenue Agropolis, BP 64501, 34394, Montpellier, Cedex 5, France
alexis.dereeper@ird.fr

[†]Contributed equally. [‡]Present address: LRSV, UMR5546 UPS, Pôle de Biotechnologies Végétales, 24 chemin de Borde Rouge, BP 42617 Auzeville, 31326, Castanet-Tolosan, France.

Keywords Genome, Annotation, System Interoperability, banana gene families

1 Introduction

The banana *(Musa acuminata)* reference genome sequence was recently released [1]. The study of the so-called DH Pahang genome has been enhanced by a number of tools and resources [2-6] that allows us mining its sequence. As a result, this whole genome sequencing project generated numerous and different types of information that we made accessible through our Bioinformatics Platform [7]. To support post-sequencing efforts, an integrative banana genome information system, articulated around an efficient Community Annotation System (CAS) was needed. For that purpose, we further developed the banana CAS and reinforced the interoperability of other databases and tools (*e.g. Galaxy, GreenPhyl, SNiPlay, pathway tools*) to propose a seamless and scalable banana genome hub.

2 Results

Our global strategy in implementing this hub was to exploit, whenever possible, generic software solutions interconnected to establish a reliable framework for scientists interested in banana and related biology. It proposes the following functionalities:

- The core of the CAS consists of three components. A relational database based on Chado [9] connected to the GBrowse genome browser [10] and the Artemis annotation editor [11].
- The Chado Controller allows access restriction, annotation quality and history [12].
- The Tripal component [13] reports features (gene). This Chado Web front end also allows interoperability with several published tools for plant genomics such as our instance of Galaxy [14] for managing sequence analysis workflows.

We presented several use cases that illustrate the use of the Banana Genome Hub [15] to retrieve data, benefit from pre-computed analyses, edit incorrect predictions and update analyses. In its current state, the Banana Genome Hub functions well to facilitate phylogenetic studies and gene family analyses.

3 Conclusion and future perspectives

The Banana Genome Hub http://banana-genome.cirad.fr/ aggregates for unified access various information systems and analytical tools that were not developed for the purpose of one specific crop. However, projects like the sequencing of the banana genome, encouraged the synergistic integration of tools. The model of CAS that we set up is a generic model applicable to other plant genomes and can thus be useful for other crop communities.

Given the development of NGS technologies, research communities face the generation of a huge volume of new data including re-sequencing of related samples, transcriptomics, transcriptional regulation profiling, epigenetic studies, high-throughput genotyping. Thus, it is important to provide a tool that centralizes, provides easy access, and allows exploiting huge amounts of data. It could sustain re-sequencing efforts and facilitate the updating of the genomic data and whole genome functional studies in *Musa*.

There will be some needs to develop additional visualization tools to highlight inter-specific synteny and structural variations between multiple Musa genomes. The transcriptomic data will require efficient tools for differential expression studies. These new data could also be useful for improving the reference genome annotation by supporting the refining of exon-intron junctions, by confirming gene expression in certain tissues and by characterizing their splice forms. Due to the importance of banana as a crop, SNP markers will need to focus our attention on the best way to represent them. Finally, data integration remains a challenge and the semantic integration of -omics data will be further investigated.

Acknowledgements

This work was supported by ANR Genoplante (grant ANR-07-GPLA-004 GNPAnnot); ANR Plant Genomics (grant ANR-08-GENM-V019 Musatract); CGIAR Research Project Root Tuber and Bananas (CRP-RTB) and by the high performance cluster of the South Green Bioinformatics Platform.

References

[1] A. D'Hont *et al.*, The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, 488:213, 2012.

[2] K. L. Howe, T. Chothia, R. Durbin, GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res*, 12:1418, 2002.

[3] T. Flutre, E. Duprat, C. Feuillet, H. Quesneville, Considering transposable element diversification in de novo annotation approaches. *PLoS One*, 6:e16526, 2011.

[4] X. Argout *et al.*, Towards the understanding of the cocoa transcriptome. *BMC Genomics*, 9:512, 2008.

[5] M. Rouard *et al.*, GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res.*, 39:D1095, 2011.

[6] P. D. Karp, S. Paley, P. Romero, The pathway tools software. *Bioinformatics*, 18:S225, 2002.

[7] South Green Bioinformatics Platform http://southgreen.cirad.fr/

[8] A. Dereeper *et al.*, SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC bioinformatics*, 12:134, 2011.

[9] C. J. Mungall, D. B. Emmert, C. FlyBase, A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, 23:i337, 2007.

[10] L. D. Stein *et al.*, The generic genome browser: a building block for a model organism system database. *Genome Res.*, 12:1599, 2002.

[11] T. Carver *et al.*, Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, 24:2672, 2008.

[12] V. Guignon *et al.*, Chado Controller: advanced annotation management with a community annotation system. *Bioinformatics*, 28:1054, 2012.

[13] S. P. Ficklin *et al.*, Tripal: a construction toolkit for online genome databases. *Database: the journal of biological databases and curation*, 2011, 2011.

[14] J. Goecks, *et al.*, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11:R86, 2010.

[15] G. Droc *et al.*, The Banana Genome Hub. *Database (Oxford)*, 2013:bat035, 2013.

OUVERTURE D'UN SERVICE D'AUTOFORMATION EN LIGNE : Pour se former en fonction de ses besoins du moment.

Sarah MAMAN[1], Christophe KLOPP[2]

[1] INRA, UMR444, Laboratoire de Génétique Cellulaire, Centre de Toulouse Auzeville, 24 Chemin de Bordé Rouge, 31320 Auzeville-Tolosane, France
sarah.maman@toulouse.inra.fr

[2] INRA, Plateforme bio-informatique Genotoul, Biométrie et Intelligence Artificielle, Castanet-Tolosan, France
klopp@toulouse.inra.fr

Keywords training e-learning NGS competencies sharing

Abstract *Une plateforme de formation en ligne (elearning) permet aux utilisateurs d'outils bioinformatiques de revenir sur des connaissances spécifiques et de ré-activer certaines compétences bioinformatiques utiles en cours de projet. Les formations en ligne reprennent le contenu des formations proposées par les différentes plateformes et donnent accès à des ressources* documentaires et pédagogiques tout en favorisant les échanges *entre apprenants. Après une étude comparative des plateformes et des outils open source disponibles, l'équipe Sigenae a choisi Dokeos pour la mise en ligne des formations en génomique sur* http://sig-learning.toulouse.inra.fr .

An elearning platform allow bioinformatics tools users to improve specific knowledge and refresh some useful bioinformatics skills during their project. Online courses incorporate the courses content offered by the different platforms and provide access to resources and exchanges between learners. Thanks to a comparative study of available open source platforms and tools , Sigenae Team choosed Dokeos for the online training in genomics on http://sig-learning.toulouse.inra.fr .

Vu le nombre exponentiellement croissant de séquences disponibles, et le nombre limité de bio-informaticiens, les biologistes sont de plus en plus nombreux a être préoccupés par le traitement de leurs fichiers. Ils participent volontiers aux formations proposées par les différentes plates-formes (liste disponible sur le site de la société française de bio-informatique http://www.sfbi.fr). Malheureusement ne se servant pas régulièrement des outils ou n'ayant pas toujours leurs données au moment de la formation, ils ne les valorisent parfois pas directement. Ces compétences spécifiques peuvent être difficiles à ré-activer plusieurs mois après la formation en salle. ils souhaitent souvent revenir sur certaines notions et astuces au cours de leur projet.

Les outils d'autoformation disponibles répondent à cette demande en reprenant dans le détail l'ensemble des éléments vus en salle. Ils peuvent ainsi refaire les exercices et les études de cas présentés, avec l'aide d'une correction interactive. A l'aide des outils d'e-learning, le biologiste a la possibilité de réactiver ses compétences au moment du besoin.

Une autoformation en ligne, ou e-learning, est une formation à distance accessible via un navigateur internet avec une interaction plus ou moins importante avec le formateur. Une formation en ligne donne accès à des ressources, des services, des échanges entre les apprenants et le formateur mais aussi entre les apprenants eux-mêmes, ouvrant ainsi des possibilités de collaboration à distance.

S'auto-former en ligne permet de se former quand on le souhaite (pas d'horaires, pas de dates fixes, pas de déplacements à prévoir , pas de contraintes logistiques) et « à la carte » en fonction des compétences que l'on cherche à approfondir. Le suivi d'une formation est ainsi allégé et moins contraignant en termes d'horaires et de temps mobilisé en salle.

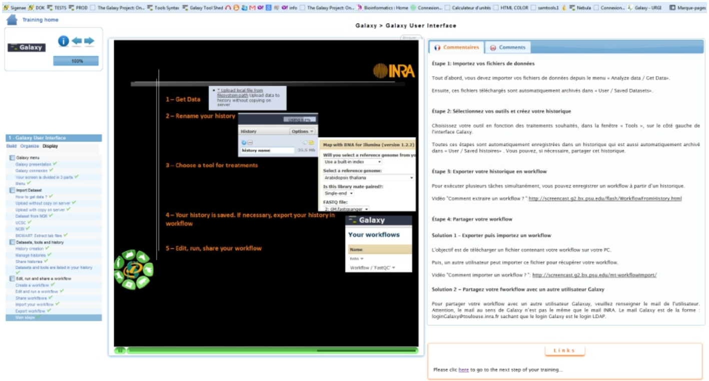
Se connecter à une plate-forme d'autoformation a plus pour objectif de transmettre et de partager des savoirs plutôt que de valider des compétences.

Il existe plus d'une dizaine de plates-formes open source de formation en ligne. Moodle et Claroline figurent parmi les plus répandues. Le dynamisme de la communauté, la simplicité et l'accessibilité de la plate-forme, tant au niveau des apprenants que des formateurs et de l'administrateur, ainsi que l'ergonomie de l'interface constituent généralement les principaux critères de choix d'une plate-forme. Mais seule une analyse fine des besoins fonctionnels et des méthodes de travail de vos équipes permet de finaliser le choix. Certaines plates-formes, comme Moodle, apportent plus d'interactivité entre les apprenants et les formateurs, et nécessitent aussi une plus forte implication des formateurs. Alors que d'autres, comme Claroline, sont plus axées sur les outils, le parcours et les méthodes pédagogiques. L'objectif d'une plate-forme est de proposer un espace de travail intuitif, avec des outils pédagogiques simples d'utilisation, un parcours pédagogique clair, des outils de communication non redondants avec les outils pré-existants de l'équipe, et enfin des supports facilement et rapidement transférable en vu d'un éventuel changement de plate-forme.

Le nombre d'outils pédagogiques open source destinés à la formation en ligne ne cesse d'augmenter. Encore une fois, des critères comme la simplicité, l'accessibilité, l'interopérabilité, et réutilisabilité nous aident à mieux évaluer et sélectionner ces outils afin d'optimiser leur intégration au parcours pédagogique.

Le poster propose une méthode d'aide au choix d'une plate-forme open source et de sélection d'outils pédagogiques. Cette méthode consiste à lister les besoins des utilisateurs (apprenants, formateurs, administrateur) afin de rechercher une adéquation avec les critères spécifiques des principales plate-formes et des principaux outils disponibles.

L'équipe Sigenae et la plate-forme Genotoul ont choisi Dokeos, de Claroline, pour dématérialiser leurs formations. La plate-forme a été installée sur une machine virtuelle CentOS (noyau Linux, base de données MySQL) accessible grâce un login et mot de passe LDAP Genotoul.


Training home
Galaxy
100%
Galaxy > Galaxy User Interface
1 - Galaxy User Interface
Build Organize Display
Galaxy menu
Galaxy presentation
Galaxy connexion
Your screen is divided in 3 parts
Menu
Import Dataset
How to get data ?
Upload without copy on server
Upload with copy on server
Dataset from NGS
UCSC
NCBI
BIOMART: Extract tab files
Datasets, tools and history
History creation
Manage histories
Share histories
Datasets and tools are listed in your history
Edit, run and share a workflow
Create a workflow
Edit and run a workflow
Share workflows
Import your workflow
Export workflow
Main steps
INRA
1 – Get Data
2 – Rename your history
3 – Choose a tool for treatments
4 – Your history is saved. If necessary, export your history in workflow
5 – Edit, run, share your workflow
History
Options
history name
Map with BWA for Illumina (version 1.2.2)
Galaxy
Your workflows
Name
Commentaires
Comments
Links

Étape 1: Importez vos fichiers de données

Tout d'abord, vous devez importer vos fichiers de données depuis le menu « Analyze data / Get Data».

Ensuite, ces fichiers téléchargés sont automatiquement archivés dans « User / Saved Datasets».

Étape 2: Sélectionnez vos outils et créez votre historique

Choisissez votre outil en fonction des traitements souhaités, dans la fenêtre « Tools », sur le côté gauche de l'interface Galaxy.

Toutes ces étapes sont automatiquement enregistrées dans un historique qui est aussi automatiquement archivé dans « User / Saved histoires» . Vous pouvez, si nécessaire, partager cet historique.

Étape 3: Exporter votre historique en workflow

Pour exécuter plusieurs tâches simultanément, vous pouvez enregistrer un workflow à partir d'un historique.

Vidéo "Comment extraire un workflow ? ":http://screencast.g2.bx.psu.edu/flash/WorkflowFromHistory.html

Étape 4: Partager votre workflow

Solution 1 - Exporter puis importez un workflow

L'objectif est de télécharger un fichier contenant votre workflow sur votre PC.

Puis, un autre utilisateur peut importer ce fichier pour récupérer votre workflow.

Vidéo "Comment importer un workflow ? ": http://screencast.g2.bx.psu.edu/mt-workflowimport/

Solution 2 – Partagez votre fworkflow avec un autre utilisateur Galaxy

Pour partager votre workflow avec un autre utilisateur Galaxuy, veuillez renseigner le mail de l'utilisateur. Attention, le mail au sens de Galaxy n'est pas le même que le mail INRA. Le mail Galaxy est de la forme : loginGalaxy@toulouse.inra.fr sachant que le login Galaxy est le login LDAP.

Please clic here to go to the next step of your training...

Voici l'adresse pour accéder aux formations en ligne : http://sig-learning.toulouse.inra.fr

Si vous n'avez pas de login Genotoul, vous pouvez vous en procurer un en demandant l'ouverture d'un compte à l'adresse suivante : http://bioinfo.genotoul.fr/index.php?id=81

Le site d'auto-formation en ligne (« sig-learning ») donne accès gratuitement et de manière illimitée, à l'ensemble des formations dispensées dans le cadre de la collaboration entre SIgenae et Bioinfo Genotoul (http://bioinfo.genotoul.fr/index.php?id=10).

Le catalogue comprend à ce jour les modules suivants :

- Unix
- cluster de calcul

- alignement génomique et recherche de variations
- alignement de transcriptome (RNA-Seq), recherche de transcrits et quantification de l'expression de gènes
- analyse d'expression des RNA non-codants (mi-RNA)

La page d’accueil de chaque formation sig-learning donne accès aux supports et aux exercices interactifs. Chaque formation est organisée en étapes et chaque étape contient plusieurs pages organisées en un menu, des diaporamas et des commentaires. La barre de progression vous indiquera à tout moment votre avancement au sein de chaque module de formation.

Sig-learning a été mise en production en octobre 2012 et compte plus de 50 utilisateurs. A chaque formation en salle, le lien vers la formation sig-learning est transmis aux stagiaires. La plate-forme bio-informatique MIGALE (http://**migale**.jouy.inra.fr/) met aussi en ligne ses modules.

L'objectif de « sig-learning » est de répondre au besoin croissant d'aide dans l'analyse des données de séquençage à haut débit, et autres domaines en fonction des besoins. Complémentaire au e-learning, le blended learning » combine l'autoformation et les formations en salle dans le but de renforcer les compétences acquises en présentiel.

Pour toute demande, veuillez envoyer un mail à sigenae-support@listes.inra.fr

Autres liens utiles:

Liste des formations génomiques	http://www.sfbi.fr
Equipe Sigenae	http://www.sigenae.org/
Demande de login BioInfo genotoul	http://bioinfo.genotoul.fr
Plate-forme e-learning MIGALE	http://migale.jouy.inra.fr
Comparateur de plates-formes	http://edutools.info
Accès sig-learning	http://sig-learning.toulouse.inra.fr
Dokeos	www.dokeos.com
Moodle	https://moodle.org/

De novo transcriptomic pipeline to provide a user friendly interface

RNAseqDenovo NGSPipelines

Philippe Bardou[2], Anis Djari[2], Claire Hoede[1], Jérôme Mariette[1], Ibouniyamine Nabihoudine[2], Céline Noirot[1], Christophe Klopp[1,2]

[1] UBIA, Plateforme Bioinformatique and [2] LGC, SIGENAE Team,
INRA, 24 Chemin de Borde Rouge – Auzeville , CS 52627 , 31326 Castanet Tolosan cedex, France
{Philippe.Bardou, Anis.Djari, Claire.Hoede, Jerome.Mariette, Ibouniyamine.Nabihoudine, Celine.Noirot, Christophe.Klopp}@toulouse.inra.fr

Keywords: RNAseq de novo, Transcriptomic data, Expression analysis, SNP detection

1 Introduction

Nowadays, sequencing technologies lead biologists to sequence more and more transcriptomes without any reference genome. After reads assembly into contigs, a huge amount of work is still required to determine the function, evaluate the level of expression, detect the SNPs and to annotate the transcript. Here we introduce (i) a pipeline to provide automatically those analyses and (ii) an extensible user interface to browse the results and perform some expression analyses.

2 Pipelinedescription

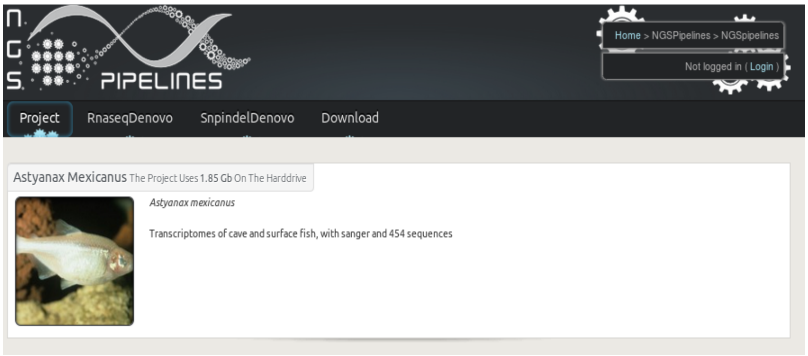
We developed a pipeline for de novo RNAseq data. From an assembly and a list of library files (fastq) the contigs are annotated with general databases (Swissprot, Refseq RNA and Refseq Protein) and species-specific databases (Ensembl, Tigr and Unigene). Thanks to the general databases annotations, we extract GO names and keywords for each contig. Contigs are then renamed by using the name of the gene corresponding to the most relevant annotation.

All libraries are mapped with bwa[1] on the contigs and expression is evaluated by counting number of reads mapped on each contig.

Variant calling is performed with samtools mpileup[2] and filtered with vcftools[3] or with the GATK package [4]. A counting of each allele per library is done. tSNPannot[5] allows us to annotate SNPs using a closed reference genome (position on this genome, consequences, distance to exon limits …).

The workflow is developed in python. Each tool is tunable by an unique configuration file.

3 Auser-friendlyinterface



Astyanax Mexicanus The Project Uses 1.85 Gb On The Harddrive

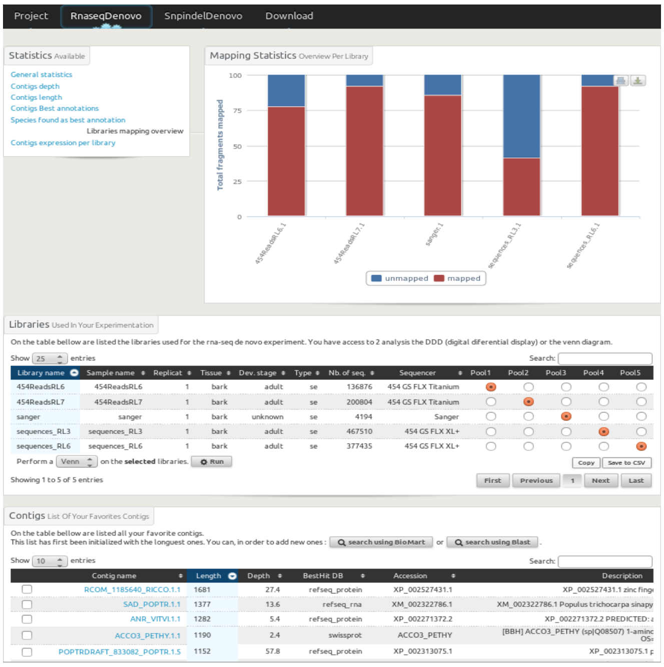
Astyanax mexicanus

Transcriptomes of cave and surface fish, with sanger and 454 sequences

Fig. 1. Overview of a project

The NGSpipelines[6] interface was developed for de novo RNAseq data. It provides several graphics and tools for exploring samples, contigs, variations and expressions data such as : (i) Venn diagram presenting samples which share contigs, (ii) Differential Digital Display supply Fisher exact test and FDR correction between two pools of samples. The user can explore his data by blast or biomart[7] forms.

This biomart plugin was designed to be easily extensible for other kind of biological applications such as rnaseq with reference genome, miRNA, BSseq… Such an interface will provide the biologists with an unique interface for all this kind of applications.



Libraries Used In Your Experimentation

On the table bellow are listed the libraries used for the rna-seq de novo experiment. You have access to 2 analysis the DDD (digital diferential display) or the venn diagram.

Show 25 entries Search:

Library name	Sample name	Replicat	Tissue	Dev. stage	Type	Nb. of seq.	Sequencer	Pool1	Pool2	Pool3	Pool4	Pool5
454ReadsRL6	454ReadsRL6	1	bark	adult	se	136876	454 GS FLX Titanium	●	○	○	○	○
454ReadsRL7	454ReadsRL7	1	bark	adult	se	200804	454 GS FLX Titanium	○	●	○	○	○
sanger	sanger	1	bark	unknown	se	4194	Sanger	○	○	●	○	○
sequences_RL3	sequences_RL3	1	bark	adult	se	467510	454 GS FLX XL+	○	○	○	●	○
sequences_RL6	sequences_RL6	1	bark	adult	se	377435	454 GS FLX XL+	○	○	○	○	●

Perform a Venn on the selected libraries. Run

Showing 1 to 5 of 5 entries

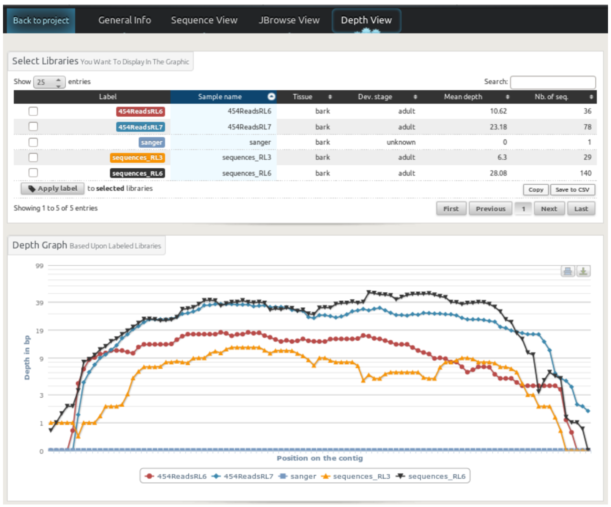
Contigs List Of Your Favorites Contigs

On the table bellow are listed all your favorite contigs.
This list has first been initialized with the longuest ones. You can, in order to add new ones : search using BioMart or search using Blast .

Show 10 entries Search:

	Contig name	Length	Depth	BestHit DB	Accession	Description
□	RCOM_1185640_RICCO.1.1	1681	27.4	refseq_protein	XP_002527431.1	XP_002527431.1 zinc fing
□	SAD_POPTR.1.1	1377	13.6	refseq_rna	XM_002322786.1	XM_002322786.1 Populus trichocarpa sinapy
□	ANR_VITVI.1.1	1282	5.4	refseq_protein	XP_002271372.2	XP_002271372.2 PREDICTED: a
□	ACCO3_PETHY.1.1	1190	2.4	swissprot	ACCO3_PETHY	[BBH] ACCO3_PETHY (sp|Q08507) 1-aminc OS=
□	POPTRDRAFT_833082_POPTR.1.5	1152	57.8	refseq_protein	XP_002313075.1	XP_002313075.1 p

Fig. 2. Overview of RnaseqDenovo



Select Libraries You Want To Display In The Graphic

Show 25 entries Search:

	Label	Sample name	Tissue	Dev. stage	Mean depth	Nb. of seq.
□	454ReadsRL6	454ReadsRL6	bark	adult	10.62	36
□	454ReadsRL7	454ReadsRL7	bark	adult	23.18	78
□	sanger	sanger	bark	unknown	0	1
□	sequences_RL3	sequences_RL3	bark	adult	6.3	29
□	sequences_RL6	sequences_RL6	bark	adult	28.08	140

Apply label to selected libraries

Showing 1 to 5 of 5 entries

Fig. 3. View of a contig depth

Availability : All the pipeline code, the interface source (developed as a plug-in for Biomart 0.8 rc6) and the documentation are available at https://mulcyber.toulouse.inra.fr/projects/ngspipelines/.

References

[1] Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 589-595.

[2] Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.

[3] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group, The Variant Call Format and VCFtools ,Bioinformatics, 2011

[4] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-303.

[5] http://mulcyber.toulouse.inra.fr/projects/tsnpannot/

[6] http://mulcyber.toulouse.inra.fr/projects/ngspipelines/

[7] Arek Kasprzyk, BioMart: driving a paradigm change in biological data management. Database (Oxford). 2011 Nov 13;2011:bar049.

Comparison of ORF prediction tools

Olivier Rué[1,2], Christophe Klopp[1,2]

[1] UBIA, Plateforme Bioinformatique and [2] LGC, SIGENAE Team,
INRA, 24 Chemin de Borde Rouge – Auzeville , CS 52627 , 31326 Castanet Tolosan cedex, France
{Olivier.Rue, Christophe.Klopp}@toulouse.inra.fr

Keywords:ORF, RNA-Seq, Benchmark

1 Introduction

Biologists studying the gene expression levels, of unsequenced organisms, de novo assemble the transcripts using RNA-Seq reads. They produce a set of contigs containing the translated regions but also surrounding UTRs and sometimes introns or intergenic regions. Coding region detection of is one of the typical tasks associated with the large-scale analysis of RNA-Seq assemblies. Some errors like frameshifts can hinder the ORF (Open Reading Frame) detection. A wrong coding region detection will induce wrong structural or functional annotations.

2 Methods

To constitute the dataset to use for this benchmark, we aligned whith fasty36 60 988 zebrafish contigs assembled with Oases (unpublished data) on 42 171 zebrafish proteins from Ensembl. We kept only perfect hits with 100% of identity.

The coordinates on the contigs of the 17 222 contigs containing a complete protein were used as reference coordinates of coding regions.

We tested 6 tools (table 1) :

- getorf [1]
- OrfPredictor [2]
- prot4EST [3]
- FrameDP [4]
- orffinder [5]
- ESTscan [6]

	getorf	orfpredictor	prot4EST	framedp	ESTscan	ORFfinder
Parameters	--find 1 --table 1	default	default	default	default	--fullstart
Use of external data	/	/	mitochondrial, ribosomal, proteic banks	/	/	/
Pre-process on EST	/	optional blastx	blastx	blastx	/	/
Method	Extracts ORF between start/stop codons	If blastx hit, used to choose the frame	Hidden Markov model, creates its own score matrices file	Hidden Markov model, self-training	Hidden Markov model, can provide a score matrices file	Extracts ORF between start/stop codons
Number of citations (June 2013)	online tool (2000)	91 (2005)	88 (2004)	31 (2009)	360 (1999)	online tool (2011)

Table1. Detail of tools used for this benchmark

For each tool, outputs were reformated to keep the prediction coordinates. Then, protein infered coordinates of ORF and predicted coordinates were compared.

The sensitivity ans specificity of predictions were computed and three classes were used to classify predictions :

- Perfect prediction : Start and Stop codons coordinates are the same as the prediction
- Shifted prediction : Prediction coordinates overlap the real coordinates but the Start and/or the Stop codons coordinate(s) is/are shifted (+- 3 nuc)
- Aberrant : Prediction coordinates do not overlap the real coordinates or the overlapping indicates that the predicted protein is not the good one (not +- 3 nuc)

3 Results

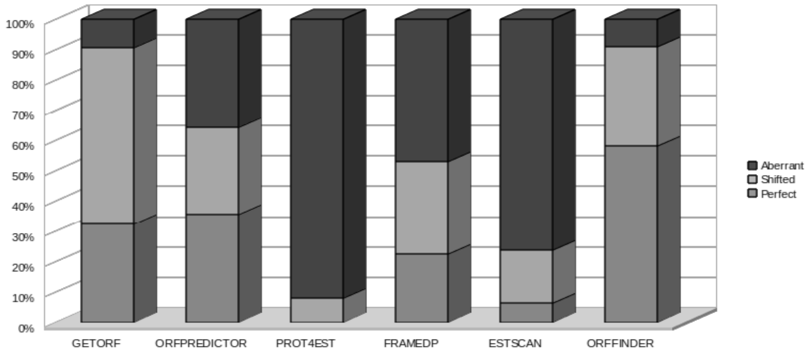


Figure1. Classification of predictions

The results showed a large disparity between tools concerning the prediction accuracy (0-58% of perfect predictions, 9-92% of aberrant predictions, Figure 1). This is due to the different strategies (Hidden Markov models, use of external banks, preliminary blasts...) and quite surprisingly the most complex ones were not the most accurate ones. For illustration, getorf [1], which extracts ORF between start and stop codons show the best results in terms of perfect predictions. Only 9% of his predictions were aberrant, whereas tools as Prot4EST [3] and ESTscan [5] showed more than 70% of aberrant predictions.

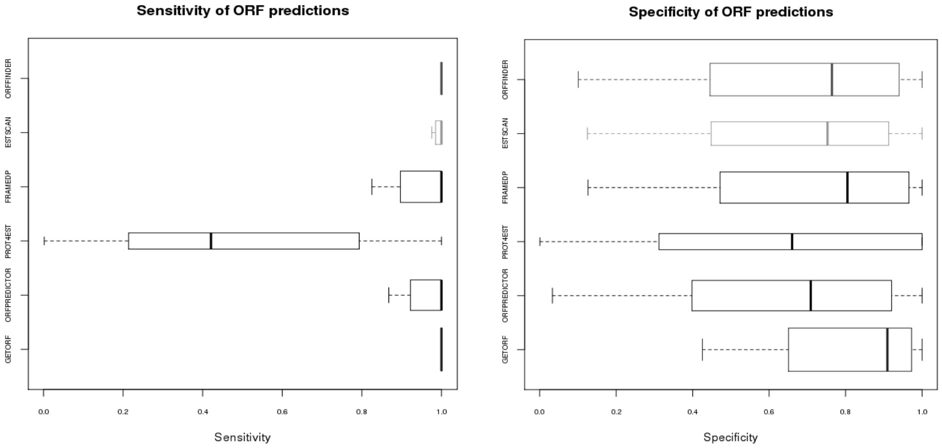


Figure2. Sensitivity and specifity of predictions

Moreover, the predictions of getorf [1] were the least distant of the real coordinates of ORF. They showed the best sensitivity (with orffinder [5]) and the best specificity (Figure 2). The sensitivity demonstrated that all predictions overlapped real proteins.

Generally, tools predict longest ORF than expected and are better to detect biologically validates CDS (mRNA).

Is the Ensembl proteins set covering all the possible proteins produced by the genome ? Publication show that genes under selection have longer 3prim UTRs which could permit to produce more alternative coding sequences [7].

In conclusion, it is very difficult to trust one tool, and it is necessary to test these tools with other data to be sure that results are reliable and are not species-specific.

References

[1] http://emboss.sourceforge.net/apps/cvs/emboss/apps/getorf.html

[2] Min, X.J., Butler, G., Storms, R. and Tsang, A. OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acid Res., 2005, Web Server Issue W677-W680 – http://proteomics.ysu.edu/tools/OrfPredictor.html

[3] Wasmuth JD & Blaxter ML: prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinformatics. 2004 Nov 30;5:187 http://www.compsysbio.org/lab/?q=prot4EST

[4] FrameDP: sensitive peptide detection on noisy matured sequences. Gouzy J, Carrere S, Schiex T. Bioinformatics. 2009 Jan 19. http://iant.toulouse.inra.fr/FrameDP

[5] http://www.ncbi.nlm.nih.gov/gorg/gorf.html

[6] ESTscan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences, Proc Int Conf Intell Syst Mol Biol. 138-48, Iseli C, Jongeneel CV, Bucher P. (1999) http://estscan.sourceforge.net/

[7] Lengthening of 3'UTR increases with morphological complexity in animal evolution, Cho-Yi Chen et al, Bioinformatics (2012) 28 (24):3178-3181

Transcriptome changes during brain aging and Alzheimer′s disease-like pathology in the cortex of the lemurian primate *Microcebus murinus*: comparison of genechips and sequencing data

RNA-seq in AD-like lemurian brains

Fabien PICHON[1], Sabine JANIN-NIDELET[2], Emeric DUBOIS[2], Jean-Michel VERDIER[1] and Gina DEVAU[1]

[1] INSERM, U710-Université Montpellier 2-EPHE, Place Eugène Bataillon, 34095, Montpellier, Cedex 5, France

[2] GENOMIX, 141 rue de la cardonille, 34094, Montpellier, Cedex 5, France

Keywords: NGS, RNA-seq, Alzheimer, aging, *Microcebus murinus*, genechips.

Aging is the major risk factor for neurodegenerative disease such as Alzheimer's disease (AD). However, the molecular events and their regulation occurring during aging still remain unclear. The aim of this study was to identify genes with expression changes in relation with age or AD in the cortex of *Microcebus murinus*, a lemurian primate, which some of them develop, as they age, the pathognomonic lesions of AD: ß-amyloid plaques and cortical atrophy (AD-like animals). We have investigated these transcriptomic changes on temporal cortex samples using human Affymetrix microarrays (HGU113 plus2). Over 14,911 transcripts detected, 152 were identified with significant changes in their expression in relation with age or with AD, and 47 revealed very discriminative [1].

These data were however incomplete because they only took into account the transcripts preserved during evolution. We have therefore pursued our investigation by high throughput sequencing (Illumina, Solexa HiSeq 2000 platform, Genomix) of 8 cortex samples (3 young adults, 3 healthy aged and 2 AD-like). A library was built for each sample, about 38 millions of short reads (50 nucleotides) was produced for each one. A preliminary analysis was performed using the Illumina’s sequencing analysis software (CASAVA 1.8.2) for assembling RNA-seq to a super set of all transcripts resulting from Ensembl, known and pseudo gene predictions (micMur 1.67). About 13,000 genes have been identified for determining differential expression. The comparison of the data of the 3 groups allowed us to detect 2,258 genes differentially expressed (edgeR software package). We compared these genes with the genes detected by microarrays. We found that 80 on the previously 152 identified genes were sorted by the two approaches. This result validated microarray analysis with human genechips. Furthermore, microarray analysis and high throughput sequencing showed complementary information. Indeed, the additional genes identified by mRNA sequencing will allow us to complete and strengthen the nature of the pathways, which essentially belong to cellular assembly and organization, and cell-to-cell signaling. Further analysis will allow us a better differentiation of the aging process from AD in the cortex.

References

[1] R. Abdel Rassoul, S. Alves, V. Pantesco, J. De Vos, B. Michel, M. Perret, N. Mestre-Francés, J.M. Verdier, and G. Devau. Distinct transcriptome expression of the temporal cortex of the primate *Microcebus murinus* during brain aging versus Alzheimer′s disease-like pathology. *PLoS One.*, 16:e12770, 2010.

Algorithmique pour l'évolution des interactions géniques

Pierre-Antoine JEAN[1], Vincent BERRY[2], Annie CHATEAU[2], Sèverine BÉRARD[3] and Eric TANNIER[4]

[1] Master STIC pour la Santé, spécialité « Bioinformatique, Connaissance, Données »
Université de Montpellier 1 & 2
Bâtiment 2, 860 rue Saint-Priest, 34090 Montpellier France
pierreantoine.jean@gmail.com

[2] LIRMM & Institut de Biologie computationnelle
UMR5506 Université de Montpellier 2, 34095 Montpellier Cedex 05 France
{vberry, annie.chateau}@lirmm.fr

[3] Centre de coopération Internationale en Recherche Agronomique pour le Développement
UMR AMAP Boulevard de la Lironde 34398 Montpellier Cedex 05 France
severine.berard@cirad.fr

[4] Centre de coopération Internationale en Recherche Agronomique pour le Développement
UMR 5558 UCB Lyon 1 – Bat Grégor Mendel 43 bd du 11 novembre 1918 69622 Villeurbanne France
eric.tannier@univ-lyon1.fr

Algorithms for the evolution of gene interactions

Résumé *La plupart des études classiques d'inférence phylogénétique considèrent l'évolution des gènes indépendamment les uns des autres. Alors que les gènes fonctionnent ensemble dans un organisme, ils interagissent, coopèrent et échangent. Il n'existe actuellement aucune méthode qui intègre dans un même modèle les évènements évolutifs qui concernent les gènes et ceux qui concernent leurs interactions. Nous explorons les problèmes algorithmiques que pose une telle intégration.*

Mots-clés Algorithmique, arbres phylogénétiques, évolution, synténies, programmation dynamique

1 Introduction

Les espèces évoluent suivant un processus arboré – appelé arbre phylogénétique : une unique espèce ancestrale (racine de l'arbre), a donné naissance à deux espèces (noeuds internes), qui se sont à leur tour divisées, et ainsi de suite jusqu'à aboutir aux espèces actuelles (feuilles de l'arbre). L'histoire évolutive des êtres vivants est la clef de la compréhension du fonctionnement des organismes que nous connaissons aujourd'hui. Les gènes portés par les espèces évoluent aussi selon des histoires qui leur sont propres, naissance, mort, duplication ou transfert d'une espèce à une autre. Si on réduit un génome à un ensemble de gènes évoluant indépendamment les uns des autres, on connaît des méthodes (modèles, algorithmes) efficaces pour proposer une histoire évolutive possible de ces gènes [1]. Mais c'est négliger que les gènes interagissent dans l'organisme et échangent de l'information. Ces relations/interactions entre gènes évoluent elles aussi, mais les méthodes qui retracent leur évolution ne sont pas aussi développées que celles retraçant l'évolution des gènes. Un algorithme permettant de retracer l'histoire évolutive d'une relation entre deux familles de gènes a été développé par [2]. Ce travail a donné lieu à un logiciel, DeCo [3], qui a été utilisé sur des données d'adjacences entre gènes pour avoir une estimation des positions relatives des gènes dans les génomes d'espèces disparues à partir des génomes contemporains. C'est cependant un travail préliminaire ne prenant en compte qu'un petit sous-ensemble des évènements évolutifs possibles. En effet les scénarios évolutifs que nous proposent DeCo ont été obtenus par une optimisation de seulement deux évènements, le gain et la cassure d'adjacence.

2 Présentation de DeCo

DeCo prend en entrée un nombre quelconque d'arbres de gènes mais les traite deux par deux, un arbre d'espèce et une liste de gènes adjacents actuels. À partir de ces données, l'algorithme réconcilie les 2 arbres de gènes (G1 et G2) avec l'arbre des espèces, sélectionne les adjacences concernant des gènes appartenant chacun à un arbre de gènes différent, puis calcule l'histoire évolutive complète de coût minimum, représentée par un ensemble d'arbres d'adjacences. Il généralise pour cela le principe de programmation dynamique de Fitch [4] pour calculer une matrice de coût minimum entre tous les couples de nœuds de G1 et G2. Puis une procédure de backtraking (top-down) s'appuyant sur cette matrice permet de construire la forêt d'arbres d'adjacences de coût minimum. Le coût de l'histoire est calculé en fonction des coûts des événements évolutifs mis en jeu et DeCo s'exécute en un temps polynomial.

3 Objectifs

Le but de notre travail est de généraliser les principes de DeCo afin de considérer des scénarios phylogénétiquement intéressants que la méthode aurait écarté. En effet, DeCo optimise seulement le nombre de cassures et de gains d'adjacences.

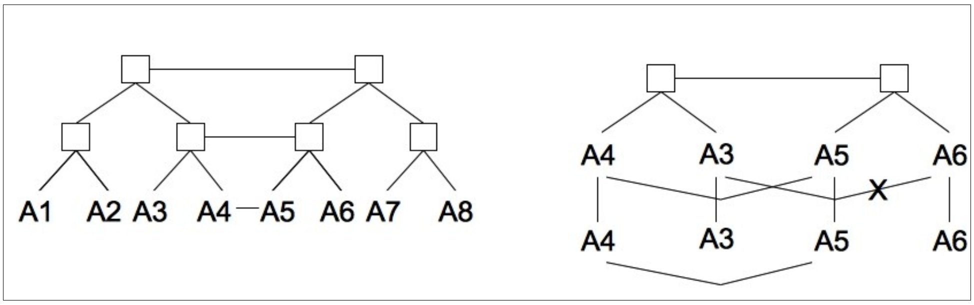


Figure 1. Le scénario de gauche explique l'adjacence A4A5 par deux duplications et le scénario de droite explique l'adjacence A4A5 par une duplication et une perte d'adjacence. DeCo favorise le scénario de gauche.

C'est pourquoi nous travaillons sur un moyen algorithmique pour compléter la fonction objectif de DeCo afin de lui intégrer d'autres évènements évolutifs importants comme les duplications, les pertes et les transferts de gènes. Pour le moment, nous avons modifié les formules de récurrences de l'algorithme pour intégrer les duplications et les pertes de gènes en bloc. Cette modification nous donne désormais un algorithme heuristique. Nous avons alors mis en place, *a posteriori* de l'algorithme deux graphes indiquants les duplications et les pertes de gènes en bloc qui se sont produits dans l'ensemble des scénarios et nous avons établi plusieurs conjectures sur la topologie de ces graphes afin de caractériser des cas où les résultats trouvés sont optimaux. Ce problème n'est pas trivial puisque l'intégration de nouveaux évènements risque de faire disparaître la propriété d'indépendance entre les évènements permettant l'utilisation des techniques de programmation dynamique. Il est important d'ajouter ces événements dans la fonction objectif, afin de reconstituer de façon fiable des génomes ancestraux, avec leurs gènes, leur structure, leurs interactions et leur fonctionnement.

Remerciements

Nous remercions le Labex NUMEV (http://www2.lirmm.fr/numev/) pour le financement de la participation à JOBIM. Ainsi que le projet ANCESTROME pour le financement de ce stage.

References

[1] GJ. Szöllősi, B. Boussau, SS. Abby, E. Tannier et V. Daubin. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc. Natl. Acad. Sci. U.S.A., vol. 109 pp.17513-8, 2012.

[2] S. Bérard, C. Gallien, B. Boussau, GJ. Szöllősi, V. Daubin et E. Tannier. Evolution of gene neighborhoods within reconciled phylogenies. Bioinformatics, 28 (18) : i382-i388, 2012.

[3] [DeCo, 2012] Logiciel DeCo programmé d'après l'algorithme de [Bérard et al, 2012]. Auteur : S. Bérard. Disponible à l'adresse http://pbil.univ-lyon1.fr/software/DeCo/

[4] WM. Fitch, Toward defining the course of evolution: minimum change for a specified tree topology. Sys. Zool., 20, 406-416.

Quantitative analysis of the repercussions of postural skull deformation on the internal skull structures using 3D imaging

Mélissa SOLINHAC[1], Guillaume CAPTIER[2] and Gérard SUBSOL[3]

[1] Master STIC pour la Santé, spécialité "Bioinformatique, Connaissances, Données"
Université de Montpellier 1 & 2, Bâtiment 2 - 860 rue Saint-Priest 34090 Montpellier France
solinhacmelissa@gmail.com

[2] CHU LAPEYRONIE 371, avenue du Doyen Gaston Giraud 34295 Montpellier CEDEX 05
g-captier@chu-montpellier.fr

[3] LIRMM de Montpellier UMR5506 CNRS/Université de Montpellier 2 34095 Montpellier CEDEX 05
gerard.subsol@lirmm.fr

Etude quantitative des répercussions des déformations posturales du crâne sur les structures internes crâniennes à l'aide d'images en trois dimensions

Résumé *La plagiocéphalie est une déformation du crâne du nouveau-né. Son étiologie reste peu connue malgré le nombre de cas qui augmentent. Certaines études tendent à montrer qu'il y a une influence sur le développement de l'enfant. Il est donc indispensable d'obtenir des données quantitatives pour essayer de classifier et d'identifier la répercussion d'une telle déformation sur les structures cérébrales. Ce travail a été mené sur des images en trois dimensions, acquises par scanner X, d'enfants atteints de plagiocéphalies.*

Mots-clés plagiocéphalie, crâne, croissance, déformation, structures cérébrales

1 Contexte

La plagiocéphalie et la brachycéphalie postérieure sans synostose sont des signes cliniques témoins d'une déformation du crâne du nourrisson. C'est une pathologie de plus en plus fréquente depuis la campagne de couchage sur le dos pour la prévention de la mort subite du nourrisson [1].

Les facteurs de risque de ces déformations sont mal connus. Le torticolis est souvent retrouvé chez les enfants atteints de plagiocéphalie [1,2,3]. La position allongée est aussi un facteur de risque important, qui peut être prévenue par des mesures posturales [4]. Les déformations seraient secondaires à des contraintes externes (traction ou compression) intervenant in utero ou dans les premiers mois de la vie [1,2]

Ces déformations posent le problème de l'évaluation de la gravité de la déformation et des répercussions neurocognitives. Actuellement l'évaluation de la gravité est basée uniquement sur des critères qualitatifs d'examen clinique [5], ce qui pose des difficultés d'évaluation de certaines thérapeutiques comme l'utilisation de casques crâniens. D'autre part, plusieurs recherches suggèrent que les enfants atteints d'une déformation auraient un risque de retard de développement [2].

2 Méthodologie

Le but est de proposer une méthode de quantification 3D de la déformation du crâne et de l'endocrâne. La quantification permettra d'avoir une échelle de gravité de la déformation, de l'asymétrie et de la répercussion sur les structures internes qui pourraient indiquer une éventuelle influence sur le développement cérébral.

Des images 3D par scanner X de cent trente-trois cas d'enfants atteints de plagiocéphalie et de brachycéphalie postérieure issus de la base de données anonymisée du CHRU de Montpellier ont été étudiées.

La segmentation est une opération de traitement d'images qui a pour but de délimiter des régions d'intérêts (ROI) dans l'image 3D. La surface du crâne et de l'endocrâne a été automatiquement segmentée à l'aide du logiciel d'analyse d'images médicales 3D Myrian® et avec le logiciel de recherche académique Endex[1] respectivement.

Le plan de symétrie sagittal médian a été déterminé automatiquement à l'aide d'un algorithme de recalage rigide. Des écarts de distances ont été calculés entre tous les points de la ROI et de son symétrique. On obtient alors une carte de dissymétrie du crâne et de l'endocrâne qui permet de localiser et de quantifier la déformation.

3 Exploitation des résultats

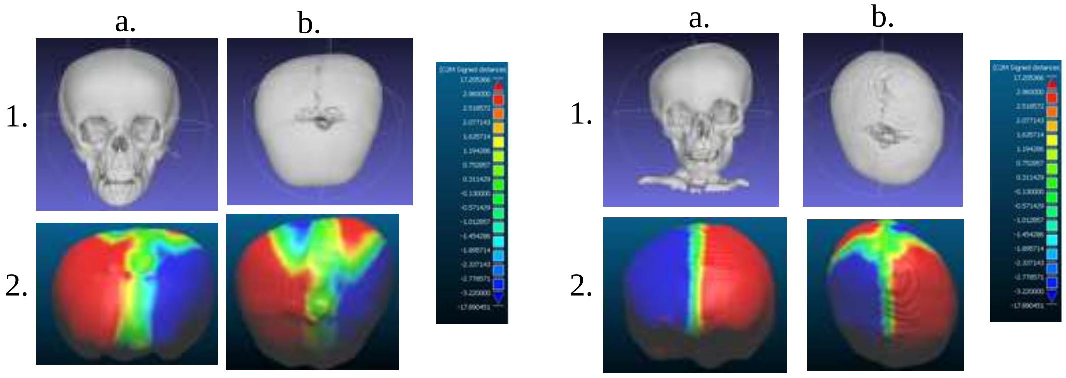


Figure 1. Deux exemples d'évaluation des déformations à l'aide des cartes de distances.
1. Visualisation surfacique du crâne, 2 Cartes des distances sur l'endocrâne ; a. Vue de face, b. Vue de dessus

Les résultats (cf. Figure 1) permettent de mettre en évidence la localisation de la déformation (frontale ou occipitale) avec les zones de méplats en bleue et les zones bombées en rouge. Les zones vertes sont symétriques. Cette visualisation permet de mieux appréhender les zones asymétriques. La carte de distance avec l'échelle permet d'avoir une quantification en mm des ces déformations. Enfin on constate que la déformation de l'endocrâne suit celle du crâne.

Cette approche va nous permettre d'étudier en particulier les déformations de la partie basale de l'endocrâne qui n'est pas accessible à l'examen clinique du crâne.

Remerciements

Nous remercions le Labex NUMEV [2] pour le financement de la participation à JOBIM 2013.

References

[1] Looman WS, Flannery AB, "Evidence-Based Care of the Child With Deformational Plagiocephaly, Part 1: Assessment and Diagnosis" Journal of Pediatric Health Care, 2012, Jul-Aug; 26(4):242-50; quiz 251-3. doi: 10.1016/j.pedhc.2011.10.003.

[2] Collett B, Breiger D, King D, Cunningham M, Speltz M, "Neurodevelpmental impliations of Deformational Plagiocephaly" J Dev Behav Pediatr. 2005 Oct; 26(5):379-89.

[3] Captier G, Leboucq N, Bigorre M, Canovas F, Bonnel F, Bonnafé A, Montoya P, Etude clinico-radiologique des déformations du crâne dans les plagiocéphalies sans synostose, Arch Pediatr. 2003 Mar;10(3):208-14

[4] Cavalier, A., M.-C. Picot, C. Artiaga, E. Mazurier, M.-O. Amilhau, E. Froye, G. Captier and J.-C. Picaud (2011). "Prevention of deformational plagiocephaly in neonates." Early Human Development 87(8): 537-543.

[5] Argenta, L. (2004). "Clinical classification of positional plagiocephaly." J Craniofac Surg 15(3): 368-372.

[1] (http://www.lsis.org/endex)
[2] (http://www.lirmm.fr/numev/)

The Effect of Distance on Hospital Choice

Auriane LANDOMIEL[1] and Nicolas MOLINARI[2]

[1] UNIVERSITE MONTPELLIER 2, place Eugène Bataillon, 34095, Montpellier, Cedex 05, France
auriane.landomiel@gmail.fr

[2] UFR DE MEDECINE, UMR729 INRA, Montpellier SupAgro, 34000, Montpellier, France
nicolas.molinari@inserm.fr

Keywords PMSI, statistics, programming, hospitalization.

1 Introduction

Les hôpitaux français remplissent une base de données de leur activité PMSI. Dans cette base, on retrouve les actes réalisés sur les patients, un numéro d'anonymat des patients, leur sexe, leur âge, et leur code géographique de résidence.

Notre équipe a développé des méthodes statistiques qui permettent de prendre en compte la distance entre les lieux d'habitation et les lieux de prise en charge des patients dans la modélisation du processus de choix d'un établissement plutôt qu'un autre.

Pour cela, le calcul des distances doit être fait. L'objet du projet/stage était donc de créer un outil qui prenne en entrée les codes géographiques de plusieurs milliers de patients inclus dans le PMSI sur une pathologie donnée, les codes postaux des hôpitaux possibles pour la pathologie donnée, et permette de sortir une matrice de distance à tous les hôpitaux de la région.

Il est ainsi nécessaire d'interfacer les tables de conversion entre les codes géographiques, les codes postaux (ou INSEE) et faire le lien avec un moteur de recherche (type Mappy) pour les calculs de distances et temps de parcours.

2 Développement

2.1 Calcul d'Itinéraires

Lors de nos recherches, nous avons étudié les fonctionnalités de plusieurs API pouvant servir à faire le type de calculs attendus : des calculs d'itinéraires. En comparant les différents sites proposant ce type d'API (GoogleMaps, ViaMichelin, Mappy, IGN Geoportail, Yahoo, Open Street Map, …), nous avons découvert que des limites étaient posées au niveau de la temporisation (requêtes espacées par des unités de temps), de la quantité (nombre de requêtes par jour) ou encore de la disponibilité (API payante au bout de plusieurs jours d'utilisation).

Nous avons décidé d'utiliser Mappy, qui offre le meilleur compromis selon nous entre ces différents critères (100 000 requêtes par jour autorisées et pas de limitation en temporisation).

2.2 Langages de Programmation

L'utilisation de l'API Mappy pour le calcul d'itinéraires induit le recours au langage JavaScript pour gérer les requêtes.

Les données en entrée du PMSI sont au format CSV, tout comme la matrice de sortie. Il a donc fallu trouver un langage permettant de manipuler facilement ce type de données. Nous avons choisi PHP, qui est apprécié pour parser des fichiers et qui permet d'interroger facilement les bases de données (*cf* 3.1). De plus, la communication entre JavaScript et PHP peut se faire facilement en utilisant Ajax.

Enfin, l'interface utilisateur est assurée par une page HTML.

3 Fonctionnalités Supplémentaires

3.1 Mise à Jour

Etant donné la masse d'informations à traiter (grand nombre de patients et d'hôpitaux), il était judicieux de mettre en place un moyen de garder en mémoire les requêtes à Mappy précédemment effectuées. Pour cela, une base de données a été créée. Elle conserve les coordonnées (latitude, longitude) correspondant aux différents codes géographiques (des domiciles et hôpitaux) ainsi que les informations relatives aux itinéraires déjà calculés (temps de parcours, distance). Ainsi, il suffit d'interroger la base et de la compléter si besoin pour générer une nouvelle matrice.

3.2 Carte Interactive

La matrice de sortie étant conséquente et sous un format textuel, nous avons pensé à mettre en place également une carte interactive à deux niveaux. Le premier représente l'ensemble des établissements concernés. Le deuxième affiche, pour chaque établissement, la localisation des domiciles des patients avec les informations relatives sur l'itinéraire correspondant.

Cette carte permet donc à l'utilisateur d'avoir un aperçu des résultats obtenus en exploitant au maximum les fonctionnalités de l'API Mappy.

4 Utilisation et conclusion

Les fichiers CSV générés par l'application sont utilisés pour des études statistiques. Nous utilisons un *modèle logit*[1] *spline*[2] pour modéliser le choix d'un établissement hospitalier pour une chirurgie du canal carpien dans le Languedoc-Roussillon. L'étude se concentre sur l'effet non linéaire de distance dans le choix d'un hôpital. Les résultats montrent que la distance possède un important et non linéaire effet sur la probabilité du choix d'un établissement. Le seuil estimé sous lequel l'effet de la distance devient négligeable est de 40 km.

L'application peut permettre d'effectuer des travaux similaires pour d'autres pathologies ou d'autres régions. La comparaison des différents résultats obtenus pourrait faire l'objet d'une étude plus approfondie de l'effet non linéaire de distance dans le choix d'un établissement hospitalier.

[1] Un *modèle logit* permet de modéliser l'effet d'un vecteur de variables aléatoires sur une variable aléatoire binomiale.

[2] Dans le domaine de l'analyse numérique, une *spline* est une fonction définie par morceaux par des polynômes.

Genetic bases of sex reversion in the African pygmy mouse Mus minutoides

A high-throughput whole-genome sequencing analysis

Thomas GUIGNARD[1], Khalid BELKHIR[2], Frédéric VEYRUNES[2] and Pierre BOURSOT[2]

[1] Master STIC pour la Santé, spécialité « Bioinformatique, Connaissance, Données »
Université de Montpellier 1 & 2, Bâtiment 2 - 860 rue Saint-Priest - 34090 Montpellier - France

[2] INSTITUT DES SCIENCES DE L'EVOLUTION CC063, UMR5554 CNRS, Place Eugène Bataillon, 34095, Montpellier, Cedex 05, France

thomas.guignard@etud.univ-montp2.fr,
{khalid.belkhir,frederic.veyrunes, pierre.boursot}@univ-montp2.fr

Keywords atypical sex determination, genomics, Next Generation Sequencing, genetics, evolution, African pygmy mouse, de novo assembly

Déterminants génétiques de la réversion de sexe chez les souris naines

Recherche par analyse de données de séquençage haut débit de génomes complets

Mots-clés Génomique, Next Generation Sequencing, détermination atypique du sexe, Souris naine, génétique, évolution, X chromosome, assemblage de novo

1 Contexte

Les systèmes de détermination du sexe sont très variés dans le monde vivant, depuis la détermination environnementale jusqu'à la détermination chromosomique en passant par la détermination génique. Comprendre l'origine de cette diversité revient en essence à comprendre comment l'évolution gère le défi de produire deux phénotypes sexuels différenciés à l'aide d'un génome unique. Quelle est la dynamique de l'interaction entre l'évolution de cet antagonisme (divergence plus ou moins grande des traits d'histoire de vie liés au sexe) et l'évolution de l'architecture génétique de la détermination du sexe (organisation génomique, type de transmission du matériel génétique d'une génération à l'autre et d'un sexe à l'autre, évolution de la recombinaison)? Les groupes d'organismes présentant des variations connues de détermination génétique du sexe entre espèces proches, ou encore mieux entre populations d'une même espèce, sont les modèles de choix pour aborder ces questions.

2 Problématique

Les souris naines africaines de l'espèce Nannomys minutoides possèdent un système de détermination du sexe particulier : deux types de chromosomes X coexistent, dont un (appelé X*) est féminisant. Ainsi seul le génotype XY est phénotypiquement mâle, les autres génotypes (XX, X*X et X*Y) étant femelles. Nous cherchons à comprendre les mécanismes évolutifs à l'origine de cette réversion de sexe chez les X*Y, et à découvrir les déterminants génétiques sous-jacents. Pour ce faire, le séquençage par une technique à haut débit (Illumina HiSeq) des génomes d'une femelle X*Y et d'une XX, ainsi que d'une femelle d'une espèce proche présentant un déterminisme du sexe « classique » ont été réalisés.

3 Approche

Les données brutes de séquençage ont permis de réaliser un assemblage de novo combiné à un mapping sur le génome connu de la souris de laboratoire. Par comparaison des séquences attribuées aux chromosomes X et X* , tout type de différence est détectable (insertions/délétions, duplications, divergence régionale de séquence, divergence fonctionnelle des gènes). Cette étude permettra de désigner des régions et gènes candidats responsables de la mutation féminisante, qui serviront de base aux analyses évolutives et fonctionnelles en aval.

Acknowledgements

This work was supported by ANR grant "SEXYMUS" (N°10-JCJC-1605), 2012 scientific grant from Fondation del Duca (Institut e France), INEE (Institut Ecologie et Environnement, CNRS and the MBB platform of the CeMEB Labex. We thank the NUMEV Labex (http://www2.lirmm.fr/numev/) for supporting the participation to JOBIM.

References

[1] Veyrunes F, Chevret P, Catalan J, Castiglia R, Watson J, Dobigny G, Robinson TJ, Britton-Davidian J, A novel sex determination system in a close relative of the house mouse. Proc R Soc B 277:1049-1056, 2010.

[2] Bachtrog D, Hom E, Wong KM, Maside X, de Jong, P, Genomic degradation of a young Y chromosome in *Drosophila miranda*. Genome Biol 9:R30, 2008.

[3] Bachtrog D, Jensen JD, Zhang Z Accelerated Adaptive Evolution on a Newly Formed X Chromosome. PLoS Biol 7:712-719, 2009.

[4] Charlesworth D, Mank J The Birds and the Bees and the Flowers and the Trees: Lessons from Genetic Mapping of Sex Determination in Plants and Animals. Genetics 186:9-31, 2010.

Differential transcriptomic analysis of target tissues during infection of *Spodoptera frugiperda* with *Junonia coenia densonucleosis* virus

Zlatomir Todorov[1,3], François Cousserans[2] , Mylène Ogliastro[2] and Thierry Dupressoir[3]

[1] Master STIC pour la Santé, spécialité « Bioinformatique, Connaissance, Données »
Université de Montpellier 1 & 2, Bâtiment 2 - 860 rue Saint-Priest - 34090 Montpellier - France
Zlatomir.Todorov@etud.univ-montp2.fr
[2] Diversité, Génomes et Interactions Microorganismes-Insectes, UMR1333 INRA CC101,
Université Montpellier 2, 34095l, Montpellier, Cedex 5, France
ogliastr@supagro.inra.fr
[3] Laboratoire de Pathologie comparée des Invertébrés, EPHE,
CC101, Université Montpellier 2, 34095l, Montpellier, Cedex 5, France
Thierry.Dupressoir@univ-montp2.fr

Abstract *This paper investigates the transcriptomic modifications induced by an insect pathogenic virus infecting the trachea cells of its host. We present the pipeline we elaborate to perform the differential analysis of our four libraries of transcriptomic data in order to exhibit candidate genes involved in the JcDNV infection of S. frugiperda.*

Keywords insect, Spodoptera, Densovirus, infection, RNA-Seq, transcriptome, analysis, differential.

1 Introduction

Viruses infecting their host *orally* are confronted to physical barriers, the epithelia, which constitute the first stage of defense against infectious agents. We have deciphered the pathophysiology of the infection of the highly polyphagous pest *Spodoptera frugiperda* (a pest noctuid of the order *Lepidoptera*) by a pathogenic *Densovirus* (DNV), the *Junonia coenia* (Jc) DNV [1] and shown that the virus first passes trough the intestinal epithelium without replication, to reach permissive cells beyond. Among the internal target tissues, other epithelia, like the trachea, the respiratory network of the larva, and the epidermis seem to be the most efficient in replicating and spreading the virus. We explored the transcriptomic response to infection in the tracheal tissues following virus entry before virus replication (1 day pi) end after (3 days pi).

2 Material and methods

In order to identify the cellular response to infection and the molecular effectors, we used the highly replicative trachea epithelium for a global analysis of the transcriptome in uninfected *vs.* infected trachea. Trachea were dissected at times 24h (before virus replication) and 72h (after virus replication) after *per os* infection of 30 4th instar larvae per time point (Ni1, Ni3, Jc1, Jc3 respectively). Messenger RNAs were isolated and sequenced using the Illumina® technology. Approximately $14x10^6$ (Ni1), $7.5x10^6$ (Ni3), $19x10^6$ (Jc1) and $24x10^6$ (Jc3) reads were obtained after adaptor cleaning and quality filtering of the libraries. We observed a distinctive pattern for the nucleotide frequencies in all four libraries ranging from the beginning of the 5’ end to the 13th nucleotide included (*FastQC*) [2]. This observation is consistent with the analysis made by Hansen *et al.* [3]. Regardless of the conclusion made by the authors that this bias “affects the uniformity of the localizations of the reads along the transcriptome”, we now assume it will not affect our differential analysis because it is equally distributed in all four libraries of RNA-Seq data. The mapping of the libraries on the reference transcriptome developed in our laboratory was performed using *Bowtie* [4] algorithm. We obtained the count data matrices using *SAMtools* [5] on the output of *Bowtie,* where only reads mapped uniquely were taken into account. Then we performed a differential expressed genes (DEG) analysis using the *DESeq* [6], *NOISeq* [7] and *edgeR* [8] packages for both time points with different comparison schema: Ni1 *vs.* Jc1 (Day_1), Ni3 *vs.* Jc3 (Day_3), Ni1 *vs.* Ni3 (Control), Jc1 *vs.* Jc3 (Infected). In *DESeq* we applied the normalization by size factor separately and in addition with a normalization by GC% and transcript length using *EDASeq* [9]. In *edgeR* we used the tagwise normalization following the *EDASeq* normalizations. With *NOISeq* we used separately three available normalization methods: *RPKM*, *UQUA*, *TMM*. The DEG candidates were processed using *Blast2GO* [10] software for an automatic functional annotation in terms of *Gene Ontology* (GO). The data bank *nr* (up to date on : May 12, 2013) from *NCBI* was used as reference for this annotation.

3 Results and Discussion

For each library, the amount of reads mapping uniquely (with one mismatch allowed) on the transcriptome is on average 47% of the library size, nearly 19% have no match regarding the criteria set to the alignment and 34% of the reads have multiple matches. After obtaining the count data we proceeded to DEG analysis. The first two comparisons Day_1 and Day_3 compare the uninfected *vs.* infected data. The Control and the Infected comparisons could indicate genes that are differentially expressed and are specific to each experimental condition. Surprisingly we observed that some transcripts were shown as DEG significant (p-value < 0.001 in *DESeq* and *edgeR*; q > 90 in *NOISeq*) by several statistical methods at the same time, but their "log2(fold change)" values calculated by the same methods were placing them opposed in terms of DEG interpretation (over-/under- expressed). Transcripts in this situation were only found by the two *DESeq* analyses, while *NOISeq* and *edgeR* were giving similar results. We named the former "ambiguous" candidates (10% of all DEG candidates) and they were discarded from the following annotation step. Thus we proceeded to the annotation of 3944 "non ambiguous" DEG candidates from all four comparisons. Finally we automatically annotated 1193 transcripts and among those, 170 were mapped to KEGG pathways.

The perspective for the bioinformatics analysis is the identification and the annotation of "stable" genes, i.e. which expression is not modified following the infection, by looking into the DEG analysis data for those with a fold change ratio proximal to 1. These genes will be used as a reference for the biological validation of the predicted differentially expressed genes by *RTqPCR*. The results of the experiment will be eventually discussed as classified for over- and under- expressed genes in condition of *Jc*DNV infection of *S. frugiperda*.

Acknowledgements

We thank the Labex NUMEV (http://www2.lirmm.fr/numev/) for funding participation at the JOBIM.conference, and Mrs. Christelle Dantec and Mr. Alban Mancheron for the technical support.

References

[1] D. Mutuel *et al*.. Pathogenesis of Junonia coenia densovirus in Spodoptera frugiperda: a route of infection that leads to hypoxia. *Virology*, 403:137-144, 2010.

[2] S. Andrews. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc

[3] K. Hansen *et al*.. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, 38:e131, 2010.

[4] B. Langmead *et al*.. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10:R25, 2009.

[5] H. Li *et al*. and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25: 2078-2079, 2009.

[6] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11: R106, 2010.

[7] S. Tarazona *et al*.. Differential expression in RNA-seq: A matter of depth. *Genome Research*, 21: 2213 - 2223, 2011.

[8] M. Robinson *et al*.. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26: 139–140, 2010.

[9] D. Risso *et al*.. GC-content normalization for RNA-Seq data. *BCM Bioinformatics*, 12: 480–507, 2011.

[10] S. Götz *et al*.. High-throughput functional annotation and data mining with the Blast2GO suite, *NAR*, 36: 3420-3435, 2008.

How UniProtKB Maps Genomes And Variants and Provide This Information.

Benoît BELY[1], Andrew NIGHTINGALE[1], Alan WILTER SOUSA DA SILVA[1] and Maria JESUS MARTIN[1]

[1] EMBL - European Bioinformatics Institute, Welcome Trust Genome Campus,CB10 1SD
Hinxton, Cambridge, United Kingdom
{bbely, anight, awilter, martin}@ebi.ac.uk

Keywords UniProtKB, complete proteome, reference proteome, protein variants, gene centric.

1 Introduction

The Universal Protein Resource (UniProt)[1], a collaboration between the European Bioinformatics Institute (EBI), the SIB Swiss Institute of Bioinformatics, and the Protein Information Resource (PIR), is comprised of four databases, each optimized for different uses. The UniProt Knowledgebase (UniProtKB) is the central access point for extensively curated protein information, including function, classification and cross-references. The UniProt Reference Clusters (UniRef) combine closely related sequences into a single record to speed up sequence similarity searches. The UniProt Archive (UniParc) is a comprehensive repository of all protein sequences, consisting only of unique identifiers and sequences. The UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for metagenomic and environmental data.

2 Providing Complete and Reference Proteomes data for complete genomes

UniProt introduces the concept of complete proteome in order to provide protein sets for complete genomes and also to re-organize proteins space[2]. A *complete proteome* consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced. A *reference proteome* is the *complete proteome* of a representative, well-studied model organism or an organism of interest for biomedical research.

Since March 2011 UniProtKB identifies and provides *complete* and r*eference proteomes* datasets. In release 2013_07 UniProtKB identified 594 *reference proteomes* and 3,361 *complete proteomes*.

2.1 How Complete proteomes are made

Complete genomes are automatically identified either from International Nucleotide Sequence Database Consortium (INSDC=EMBL/GeneBank/DDBJ) and Ensembl or EnsemblGenomes (EnsemblBacteria, EnsemblFungi, EnsemblMetazoa, EnsemblProtist and EnsemblPlant). UniProt curators review and validate the complete proteomes, they select reference proteomes according to user community. Then all proteins annotated in underline nucleotide sequence will be tagged with “complete proteome” keyword and complemented with publication of sequencing genome project. Therefore a protein translated from cDNA (mRNA) might not be flagged to be part of complete proteome unless it is 100% full length identical to an other protein already flagged.

2.2 How Complete proteomes data can be retrieved

Proteome datasets can be download using UniProtKB tools REST (1), UniProtJAPI (2) or UniProtBioMart (3). Additionally for 12 species, including the 8 previously provided by IPI and complemented with C.elegans, Dog, D. melanogaster and Pig, UniProt make fasta files available on FTP site (4), it is the continuation of FTP downloads for species provided by IPI.

For example the following URL allow you to retrieve all complete proteome for Human in fasta format including splice variant isoforms :

http://www.uniprot.org/uniprot/?query=organism:9606+keyword:181&format=fasta&include=yes

Accessing underlining nucleotide sequence can be done by using idmapping tool (2). After retrieving UniProtKB accessions list from specified set (eg: Human complete proteome), idmapping tool can be queried to retrieve INSDC accessions.

(1) http://www.uniprot.org/faq/28
(2) http://www.ebi.ac.uk/uniprot/remotingAPI
(3) http://www.ebi.ac.uk/uniprot/biomart/martview/
(4) ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomes/

3 Usage of complete proteomes

3.1 Quest for Orthologs (QfO)

Since 2009 UniProt team is working in collaboration with QfO[3] consortium in order to provide a gene centric benchmark data set for complete proteomes. This dataset is released once a year and concern 147 species for 2013 base on UniprotKB release 2013_04. This benchmark provide to the ortholog community a standard data set to effectively compare their methods. For each of the proteomes we provide : one fasta file containing non-redundant FASTA sets for the canonical sequences, gene to protein mapping file and idmapping containing all cross-references link to those proteins.

Data available from http://www.ebi.ac.uk/reference_proteomes/.

3.2 Non-synonymous variation data

UniProt has annotated the complete Homo sapiens proteome and approximately 20,000 protein coding genes are represented by a canonical protein sequence in UniProtKB/Swiss-Prot. Most of these protein sequences are now mapped to the reference genome assembly produced by the international Genome Reference Consortium (GRC). This mapping was possible through a process of aligning all human protein sequences in the UniProtKB to the protein translations in Ensembl, based on 100% amino acid identity over the entire sequence. Where protein sequences were not found in UniProtKB, new records were added to the UniProtKB reference proteome data set.

Studies on sequence variation are increasing; projects like 1000 Genomes and the Cancer Genome Project are generating a vast amount of variant information that is, or will be, stored by Ensembl variation in their databases. This exponential growth of variation information is making it infeasible to expect non-synonymous variants to be only manually assessed and added to UniProtKB as it is done currently. With human reference protein sequences mapped to the reference genome assembly, UniProt has developed a pipeline to import high-quality 1000 Genomes[4] and COSMIC[5] *missence variants*[1] from Ensembl variation. 389,935 single amino acid variants have been identified for import in the UniProt human reference proteome. The variation data provides to clinical medicine and biological science invaluable information. UniProt intends to extend this Ensembl variant import pipeline for other species with a complete proteome.

References

[1] R. Apweiler, A. Bairoch, CH. Wu, WC. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, MJ. Martin, DA. Natale, C. O'Donovan, N. Redaschi and LS. Yeh. UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res.,* 32:D115-9, 2004.

[2] UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*., 40:D71-5, 2012.

[3] T. Gabaldon, C. Dessimoz, J. Huxley-Jones, A. Vilella, E. Sonnhammer and S. Lewis. Joining forces in the quest for orthologs. *Genome Biology*, 9:403, 2009.

[4] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 467 (7319): 1061–1073, 2010.

[5] SA. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet,* Chapter 10: Unit 10-11, 2008.

[1] A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved. (http://www.sequenceontology.org/browser/current_release/term/SO:0001583)

The *Aphanomyces euteiches* Genome Project: first genomic insights into the *Aphanomyces* genus, encompassing plant and animal pathogens

Juliette Lengellé[1], Hélène San Clémente[1], Christophe Klopp[2], Patrick Wincker[3], Bernard Dumas[1] and Elodie Gaulin[1]

[1] LRSV, UMR5546 UPS/CNRS, 24, chemin de Borde Rouge, Pôle de Biothechnologie Végétales, 31326, Castanet-Tolosan, B.P.42617 Auzeville, cedex, France
{lengelle, sancle, dumas, gaulin}@lrsv.ups-tlse.fr

[2] INRA, MIA, 24, chemin de Borde Rouge, Pôle de Biothechnologie Végétales, 31326, Castanet-Tolosan, B.P.42617 Auzeville, cedex, France
Christophe.Klopp@toulouse.inra.fr

[3] Genoscope, CEA, 2, rue Gaston Crémeiux, Centre National de Séquençage Institut de génomique Direction des Sciences du vivant, 91057, Evry, CP5706, cedex, France
pwincker@genoscope.cns.fr

Keywords oomycete, *Aphanomyces*, genome, analyses, annotation, comparative genomics, effectors, high-throughput sequencing

Oomycetes are filamentous eucaryotic organisms, phylogenetically distant from ‘true fungi’ and related to brown algae and diatoms. These microorganisms are responsible of many animal and plant diseases. Oomycetes are classified in two orders: *Peronosporales* encompassing almost exclusively phytopathogenic species for which numerous complete genome sequences are available and *Saprolegniales* for which few genomic informations are yet available [1, 2]. Interestingly, the *Aphanomyces genus (Saprolegniale)* is the only oomycete genus including zoo, phyto and non pathogenic species. The first complete genome of *Aphanomyces euteiches [ATCC 201684]*, a root parasite of legumes [1] was sequenced in collaboration with the CNS-Genoscope and the Bioinformatic plateform of Toulouse. To determine the molecular mechanism underlying host adaptation, eight genomes of *Aphanomyces* representing the biological diversity in the genus, are currently sequenced by Genomic plateform of Toulouse.

De novo assembly of *Aphanomyces euteiches* 454 data has allowed to obtain 17X coverage, with a genome size of 61Mb. This genome had a GC content around 51% and 23% of repeat regions. 90% of RNASeq sequences have been mapped on the assembly. More 19, 000 genes were predicted by Augustus for which 10% encode for secreted proteins (SignalP). The genome characteristics of *Aphanomyces euteiches* are similar of those observed on the fish pathogen *Saprolegnia parasitica* [3].

The infection success of pathogenic microorganism has depended of secreted proteins called effectors, which were aided parasite to infect his host. A research of these effectors in *Aphanomyces euteiches* has revealed a large panel of proteases and kinases compared as other oomycetes. Surprisingly, *Aphanomyces euteiches* hadn't RxLR genes, known to be involved in oomycetes pathogenicity and widely distributed in *Phytophtora infestans* (>500 genes), the causal agent of late blight of potatoes [2]. In contrast, *Aphanomyces euteiches* had around 120 Crinkler's genes (CRNs) which were firstly characterized in *Phytophtora infestans* (>300 genes). CRNs are preferentially organized as cluster and located nearby repeat regions in *Aphanomyces euteiches* genome. Some CRNs family are specific to *Aphanomyces e*uteiches, suggesting that these proteins had a particular role during the host interaction. The last data regarding *Aphanomyces euteiches* secretom analyses and the data generated on the eight other *Aphanomyces* species will be presented.

Acknowledgements

This work was supported by CNS-Genoscope (AP09/10) and ANR JCJC (ANR-12-JSV6-0004-01).

References

[1] E. Gaulin, C. Jacquet, A. Bottin, B. Gaulin. Root rot disease of legumes caused by Aphanomyces euteiches. *Mol Plant Pathol*., 8(5):539-548, 2007.

[2] B.J.Hass et al. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature*, 461:393-398, 2009.

[3] *Saprolegnia genome* Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/)

Analysis and modeling of chromosome dynamics during mitosis in fission yeast

Hadrien MARY[1], Guillaume GAY[2], Clémence GRUGET[1], Sylvie TOURNIER[1] and Yannick GACHET[1]

[1] LBCMCP, UMR5088 CNRS, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France
hadrien.mary@univ-tlse3.fr
[2] mitotic-machine.org, Analyse de données et modélisation pour la biologie
gllm.gay@gmail.com

Keywords imaging, mitosis, chromosome segregation, *S. pombe*, kinetochores, microtubules, segmentation, tracking, modeling, video microscopy

Mitosis is a highly preserved process in all eukaryotic cells. A crucial step during chromosomes segregation is the transition between metaphase to anaphase, where sister chromosomes segregate and migrate to their respective cell poles. Any defect at this stage can lead to abnormal number of chromosomes in daughter cells called aneuploidy.

The segregation of the genetic material is achieved by a specialized structure called the mitotic spindle [1], composed of kinetochores (KT), KT-associated motors (kinesin 8, dynein, CENP-E) and regulators of microtubule (MT) dynamics (dam 1, kinesin 13, Aurora B kinase). During metaphase, sister KT pairs oscillate along mitotic spindle before to align at the cell center to form the mitotic plate.

We use *Schizosaccharomyces pombe* (*S. pombe*) as a model system to study the control of the eukaryotic cell cycle. With only three chromosomes, *S. pombe* is the perfect candidate to analyze chromosome dynamics.

In order to define and characterize key protein functions, we are currently setting up a robust analysis workflow from video microscopy acquisition to trajectories analysis. We are implementing already published tools such as peak detection algorithm [2] and a robust single particle tracker [3]. We also take advantage of modern mathematical tools to automatically process results from tracking such as supervised machine learning implementation to be able to detect anaphase from trajectories. The workflow uses common and widely spread Python scientific libraries such as numpy, scipy, matplotlib, pandas, scikits-learn and scikits-image. In a recent study [4] our team demonstrated that a simple model could describe chromosome segregation, with a focus on the correction of the errors in the attachment of the microtubules to the kinetochore. A particular effort was made on reducing the number of parameters and having them based on actual *in vivo* data.

Using these tools, we demonstrate that the mechanisms controlling kinetochore oscillations can be separated from the process of chromosome alignment at the metaphase plate. We identify the key proteins involved in these two processes, namely the motor protein, kinesin 8 and the microtubule associated protein, Dam 1.

Acknowledgments

This work was supported by the ANR Blanc 1206 01 “Chromocatch”and the FRM (ING 2012 1226 280) “innovative project”.

References

[1] J Richard Mcintosh, Kirill Zhudenkov, Mary Morphew, Cindi Schwartz, Fazly I Ataullakhanov, and Ekaterina L Grishchuk. “Conserved and divergent features of kinetochores and spindle microtubule ends from five species”. In: *The Journal of Cell Biology* 200.4 (2013), pp. 459–474.

[2] Arnauld Sergé, Nicolas Bertaux, Hervé Rigneault, and Didier Marguet. “Dynamic multiple-target tracing to probe spatiotemporal cartography of cell membranes.” In: *Nature methods* 5.8 (Aug. 2008), pp. 687–94.

[3] Khuloud Jaqaman, Dinah Loerke, Marcel Mettlen, Hirotaka Kuwata, Sergio Grinstein, Sandra L Schmid, and Gaudenz Danuser. “Robust single-particle tracking in live-cell time-lapse sequences.” In: *Nature methods* 5.8 (Aug. 2008), pp. 695–702.

[4] Guillaume Gay, Thibault Courtheoux, Céline Reyes, Sylvie Tournier, and Yannick Gachet. “A stochastic model of kinetochore-microtubule attachment accurately describes fission yeast chromosome segregation.” In: *The Journal of cell biology* 196.6 (Mar. 2012), pp. 757–74.

Integrated next generation sequencing storage and processing environment

Jérôme Mariette[1], Frédéric Escudié[2], Antoine Leleu[1], Philippe Bardou[3], Claire Kuchly[2], Gérald Salin[2], Christophe Djemiel[1], Claire Hoede[1], Céline Noirot[1,] Christophe Klopp[1,2]

[1] UBIA, Bioinformatic platform, [3] LGC, Genomc platform and [3] LGC, SIGENAE,
INRA, 24 Chemin de Borde Rouge – Auzeville , CS 52627 , 31326 Castanet Tolosan cedex, France
{Jerome.Mariette, Frederic.Escudie, Antoine.Leleu, Philippe.Bardou, Claire.Kuchly, Gerald.Salin, Christophe.Djemiel, Claire.Hoede, Celine.Noirot, Christophe.Klopp}@toulouse.inra.fr

Keywords: Storage, Workflows, Next Generation Sequencing

1 Background

Next generation sequencing platforms are now well implanted in sequencing centres and some laboratories. Upcoming smaller scale machines such as the 454 junior from Roche or the MiSeq from Illumina will increase the number of laboratories hosting a sequencer. In such a context, it is important to provide these teams with an easily manageable environment to store and process the produced reads.

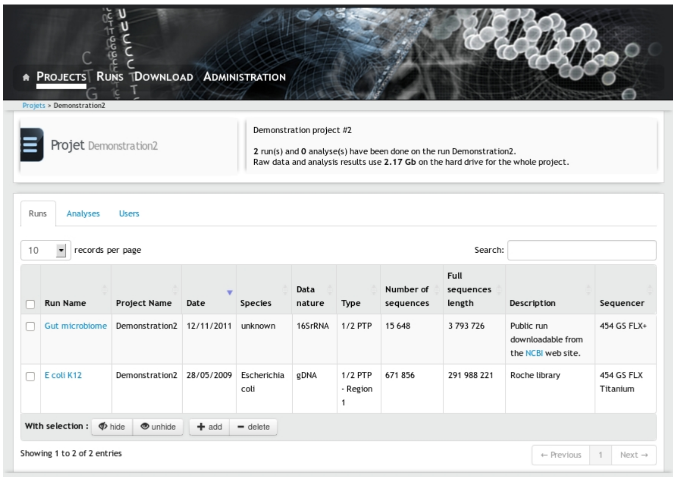
NG6 [1] is an information system providing a set of automated analysis pipelines built to process NGS (Next Generation Sequencing) data which can be executed locally or in a cluster environment. The v2.0 is built upon the jflow [2] workflow management system.

2 Implementation

NG6 offers three user statuses associated to a run (administrator, manager and member) and two data access levels: public and private.

If the connected user is a project administrator, he will be able to run available workflows within the graphical user interface. Once the pipeline execution is over, all newly added analysis and runs will be flagged as hidden. This is meant to permit the validation of the run by the team in charge of the sequencer before data release to manager. When the data are released, the manager is then allowed to give access to project members.

The connected user will be able to download raw and processed data and browse through the analysis results, statistics and parameters.



Demonstration project #2

2 run(s) and **0** analyse(s) have been done on the run Demonstration2.
Raw data and analysis results use **2.17 Gb** on the hard drive for the whole project.

10 records per page

□	Run Name	Project Name	Date	Species	Data nature	Type	Number of sequences	Full sequences length	Description	Sequencer
□	Gut microbiome	Demonstration2	12/11/2011	unknown	16SrRNA	1/2 PTP	15 648	3 793 726	Public run downloadable from the NCBI web site.	454 GS FLX+
□	E coli K12	Demonstration2	28/05/2009	Escherichia coli	gDNA	1/2 PTP - Region 1	671 856	291 988 221	Roche library	454 GS FLX Titanium

With selection :

Showing 1 to 2 of 2 entries

Fig.1.: Administrator view of a run

NG6 includes a workflow environment already containing pipelines adapted to different input formats (sff, fasta, fastq and qseq), different sequencers (roche 454, Illumina hiseq and miseq) and various analyses : quality check (casava, illumina, roche 454), rnaseq (hiseq) and diversity (miseq, roche 454).

3 Extending NG6

Adding new analysis component into NG6 requires two steps. The first one is writing the jflow component of the analysis using the NG6 API to define the data stored in the database and the result files stored in the directory structure. Second, a smarty [3] template is specified to set the corresponding analysis display. While writing the smarty template, the developer has access to several objects to build the analysis display as wished.

Moreover, NG6 can easily be extended to handle data outside of the NGS world such as metabolomic data which should be the next kind of data handled by the system.

References

[1] Mariette J, Escudie F, Allias N, Salin G, Noirot C, Thomas S, Klopp C. NG6: Integrated next generation sequencing storage and processing environment. BMC Genomics 2012, 13:462.

[2] https://mulcyber.toulouse.inra.fr/plugins/mediawiki/wiki/jflow/index.php/Accueil

[3] http://www.smarty.net/

Jflow: A fully scalable Javascript workflow management system

Jérôme Mariette[1], Ibouniyamine Nabihoudine[2], Philippe Bardou[2], Christophe Klopp[1,2]

[1] UBIA, Bioinformatic platform, and [2] LGC, SIGENAE,
INRA, 24 Chemin de Borde Rouge – Auzeville , CS 52627 , 31326 Castanet Tolosan cedex, France
{Jerome.Mariette, Ibouniyamine Nabihoudine, Philippe.Bardou,
Christophe.Klopp}@toulouse.inra.fr

Keywords: Workflows, Next Generation Sequencing

1 Background

Workflow management systems (WMS) are defined as software packages managing and executing computational pipelines. Nowadays, such systems are widely used in bioinformatics because they enable researchers to analyse the large amount of data generated by high throughput platforms.

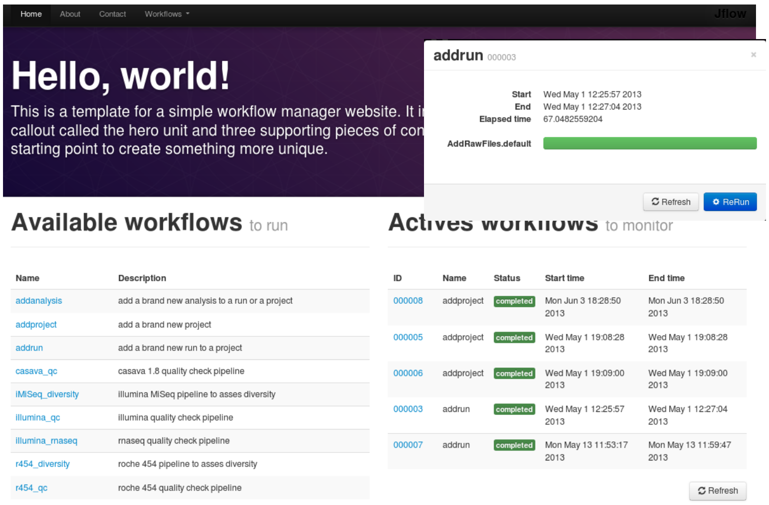
Some WMS like Galaxy [1] enable users to process their data using collection of local tools through web forms. Galaxy is probably the most used of such systems because it is based on a comprehensive web interface designed for tool and database integration. BioMOBY [2] and Taverna [3] differ from others WMS because they organize and integrate multiple web service providers. The main limit of such WMS is the network performance, service availability and I/O compatibility between providers.

All these WMS have their own user interface and can hardly be used as components of an existing web project. Nowadays, in order to provide access to new tools, it is quite common to implement a web portal that wraps the execution of the given software package. We describe jflow which includes on one hand the core of the WMS able to execute workflows defined by a set of components, and on the other hand a fully scalable jquery [4] plugin able to request the jflow REST API.

2 Implementation

Jflow core is based upon the Makeflow [5] workflow engine and weaver [6], its Python API. Adding a new component in the system requires to write a Python class inheriting from the "Component" class and to overwrite the "process" method wrapping the new tool. In the same way, writing a workflows consist of inheriting from the "Workflow" class and overwriting the "process" method. In this last method, the workflow is defined as the succession of components. Moreover, a property file should be created to define the workflow parameters.

All information about a workflows will be accessible from both jflow command line interface and its REST API. Thus, users can list available workflows and their states, run and monitor them. Accessing those functionalities from the command line interface can easily be done using the jflow_cli.py command. The same thing can also be done from a website integrating the jflow plug-in. To do so, jflow offers four modules to retrieve workflows information.



Hello, world!

This is a template for a simple workflow manager website. It i
callout called the hero unit and three supporting pieces of con
starting point to create something more unique.

addrun 000003

Start Wed May 1 12:25:57 2013
End Wed May 1 12:27:04 2013
Elapsed time 67.0482559204

AddRawFiles.default

Available workflows to run

Name	Description
addanalysis	add a brand new analysis to a run or a project
addproject	add a brand new project
addrun	add a brand new run to a project
casava_qc	casava 1.8 quality check pipeline
iMiSeq_diversity	illumina MiSeq pipeline to asses diversity
illumina_qc	illumina quality check pipeline
illumina_rnaseq	rnaseq quality check pipeline
r454_diversity	roche 454 pipeline to asses diversity
r454_qc	roche 454 quality check pipeline

Actives workflows to monitor

ID	Name	Status	Start time	End time
000008	addproject	completed	Mon Jun 3 18:28:50 2013	Mon Jun 3 18:28:50 2013
000005	addproject	completed	Wed May 1 19:08:28 2013	Wed May 1 19:08:28 2013
000006	addproject	completed	Wed May 1 19:09:00 2013	Wed May 1 19:09:00 2013
000003	addrun	completed	Wed May 1 12:25:57 2013	Wed May 1 12:27:04 2013
000007	addrun	completed	Mon May 13 11:53:17 2013	Mon May 13 11:59:47 2013

Fig.1.: An example of jflow integration

These modules are fully scalable and provide multiple methods and events to ease jflow integration. As example, the “click” event on a specific workflow triggers an event that can be listen to and used to build a workflow form wherever the web site designer wants it to be showed.

In order to ease jflow component creation a Galaxy configuration file parser will be provided. It will enable to easily provide jflow access to Galaxy components.

References

[1] Giardine B, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15:1451-1455.

[2] Wilkinson MD, Links M. BioMOBY: an open source biological web services proposal. Brief Bioinform. 2002;3:331-341.

[3] Oinn T, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 2004;20:3045-305

[4] http://jquery.com/

[5] Michael Albrecht, Patrick Donnelly, Peter Bui, and Douglas Thain, Makeflow: A Portable Abstraction for Data Intensive Computing on Clusters, Clouds, and Grids,Workshop on Scalable Workflow Enactment Engines and Technologies (SWEET) at ACM SIGMOD, May, 2012.

[6] Peter Bui, Compiler Toolchain For Data Intensive Scientific Workflows, Ph.D. Thesis, University of Notre Dame, June, 2012.