# Assessing the enrichment significance of a possition weight matrix (PWM) along a DNA sequence

Julien J. Dumazert, Jean-Yves J.-Y. Stephan, Marie Agnes M. A. Petit, Sophie S. Schbath

# JOBIM

JOURNÉES
OUVERTES
BIOLOGIE
INFORMATIQUE
MATHÉMATIQUES

# TOULOUSE 2013

Volume 1/2 **PRÉSENTATIONS ORALES**

[ 1-4 JUILLET 2013 ]


# **J**ournées
# **O**uvertes
# **B**iologie
# **I**nformatique
# **M**athématiques

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
Jérome 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 ordre de présentation)

# Session 1 : Analyse de séquences

## Conférence invitée

SIMON ANDERS

European Molecular Biology Laboratory (EMBL)
Heidelberg, Germany

### Signal and noise in quantitative RNA sequencing

A standard application of RNA-Seq is the comparisons of transcriptomes; in the simplest case, finding genes whose expression differs between, say, treated and control samples. Usually, the goal is to estimate the average change in expression for each gene. However, to get reliable conclusions and guard against false positives, it is essential to also assess how precise these estimates are. Quantifying this precision, rather than estimating the change itself, is the difficult part in developing analysis methods. From this viewpoint, I will review the fundamental concepts in quantitative data analysis for RNA-Seq, discuss the current state of the art and the still open questions, and suggest some new conceptual approaches. I will also present novel aspects of our own work on the topic, including the differences between our popular "DESeq" tool and our new method, "DESeq2".

# CRAC, a software for analyzing RNA-seq reads

Nicolas Philippe[1,2,4*], Mikaël Salson[3*], Thérèse Commes[2,4] and Eric Rivals[1,4]

[1] LIRMM, UMR 5506, CNRS and Université de Montpellier 2, Montpellier, France
{nicolas.philippe,rivals}@lirmm.fr
[2] IRB, U1040 INSERM, Montpellier, France
{nicolas.philippe,therese.commes}@inserm.fr
[3] LIFL (UMR CNRS 8022, University of Lille) and Inria Lille-Nord Europe, France
mikael.salson@lifl.fr
[4] Institut de Biologie Computationnelle, Montpellier, France

**Abstract** *A large number of RNA-sequencing studies set out to predict mutations, splice junctions or fusion RNAs. We propose a method, CRAC, that integrates genomic locations and local coverage to enable such predictions to be made directly from RNA-seq read analysis. A k-mer profiling approach detects candidate mutations, indels, splice or chimeric junctions in each single read. CRAC increases precision compared with existing tools, reaching* 99.5% *for splice junctions, without losing in sensitivity. Importantly, CRAC predictions improve with read length. In cancer libraries, CRAC recovered* 74% *of validated fusion RNAs and predicted novel recurrent chimeric junctions. CRAC is available at* http://crac.gforge.inria.fr.

**Keywords** RNA-seq, high-throughput sequence analysis, mapping, splicing,fusion transcript

## 1 Introduction

As High-Throughput Sequencing (HTS) improves and goes cheaper, bioinformatic analyses become more critical and time consuming. They still follow the same paradigm as in the first day of HTS technologies: a multiple step workflow – mapping, coverage computation, and inference – where each step is heuristic, concerned with only a part of the necessary information, and is optimized independently from the others. Consequently analyses suffer from the drawbacks inherent to this paradigm: (a) pervading erroneous information, (b) lack of integration, and (c) information loss, which induces re-computation at subsequent steps and somehow prevents cross-verification. For instance (b), the mapping step cannot use coverage information, which prevents it from distinguishing biological mutations from sequencing errors early in the analysis.

Here, we design a novel and integrated strategy to analyze reads when a reference genome is available. Our approach extracts the information solely from the genome and read sequences, and is independent of any annotation; we implemented it in a program named CRAC. The rationale behind it is that an integrated analysis avoids re-computation, minimizes false inferences, and provides precise information on the biological events carried by a read. A peculiarity of CRAC is to deliver computational predictions for point mutations, indels, sequence errors, normal and chimeric splice junctions, in a single run. CRAC is compared with state-of-the-art tools for mapping (BWA/SOAP2/Bowtie/GASSST) [5–8, 12], and both normal (GSNAP/TopHat/MapSplice) [13–15] and chimeric (TopHat-fusion) [4] splice junction predictions. The results show the relevance of the approach in terms of efficiency, sensitivity, and precision (which is also termed specificity in the literature). We also provide true assessments of all methods sensitivity by analyzing complex simulated data.

Availability: CRAC is distributed under the GPL-compliant CeCILL-V2 license and is available as a source code archive or a ready to install Linux package from the CRAC project website http://crac.gforge.inria.fr/ or the ATGC bioinformatics platform http://www.atgc-montpellier.fr/crac. It includes two programs: `crac-index` to generate the index of the genome, and `crac` for analyzing the reads. A full version of this paper has been published in *Genome Biology* [10].

* NP and MP equally contributed to this work

# 2 Algorithm

## 2.1 Overview

CRAC is a method to analyze reads when a reference genome is available, although some procedures can be useful in other contexts. It uses only the read collection and the reference genome. It is thus independent of annotations. Here, analyzing reads is detecting biological *events* (*i.e.* mutations, splice junctions, chimeric RNAs) and sequencing errors from RNA-seq reads.

CRAC analysis is based on two basic properties (P1-P2).
**P1** For a given genome length, a certain sequence length $k$ is sufficient to match in average to a unique genomic position with high probability. $k$ can be computed and optimized [9]. Thus, in a read any $k$-*mer* (a $k$-long substring) is used as a witness of a possible location in the genome. In average over all $k$-mers, the probability of getting a *false location (FL)* is $\simeq 10^{-4}$, with $k = 22$, for the Human genome size [9].
**P2** As reads are sequences randomly sampled from biological molecules, several reads usually overlap a range of positions from the same molecule. Hence, a sequencing error, contrary to a biological variation, occurring in a read should not affect the other reads covering the same range of positions.

CRAC proceeds each read, of length $m$, in turn. It considers the $k$-mers starting at any position in the read (*i.e.* $m - k + 1$ possible $k$-mers). It computes two distinct $k$-mer profiles:

**the location profile** records for each $k$-mer its exact matching locations on the genome and their number,

**the support profile** registers for each $k$-mer its *support*, which we define as the number of reads where this $k$-mer occurs. The support value has a minimum value of one since the $k$-mer exists in the current read.

CRAC analyzes jointly these two profiles to detect multiple situations and predicts in a single analysis sequencing errors, as well as potential genetic variations, splice junctions, or chimeras. The location profile is computed using a compressed index of the reference genome, such as a compressed Burrows Wheeler Transform [2], while the support profile is obtained on-the-fly by interrogating a specialized read index (Gk-arrays [11]). CRAC uses pairing information of paired end reads as a confirmation for the prediction of chimera.

Clearly, the support is a proxy of the coverage and allows exploiting property P2 for distinguishing sequencing errors from variations, and gaining confidence in predictions. As illustrated below, the location profile delivers a wealth of information on the mapping situation, but detecting the concordance of variations in the two profiles makes the originality of CRAC.

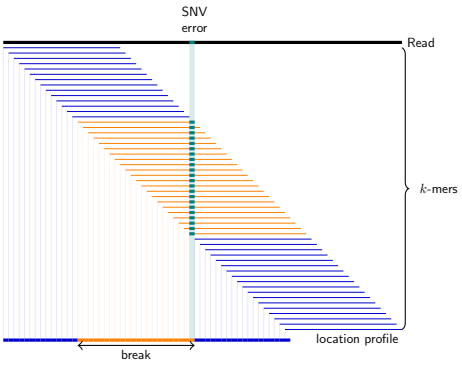
## 2.2 Algorithm description

In a collection, some reads will exactly match the reference genome, while others will be affected by one or more differences. Here, we describe how a read is processed and concentrate on reads that differ from the reference. For clarity of exposure, we make simplifying assumptions: a/ $k$-mers have no false genomic locations [1], b/ the read is affected by a single difference (substitution, indel, splice junction), c/ this difference is located $> k$ nucleotides away from the read's extremities (otherwise, we say it is a *border case*).

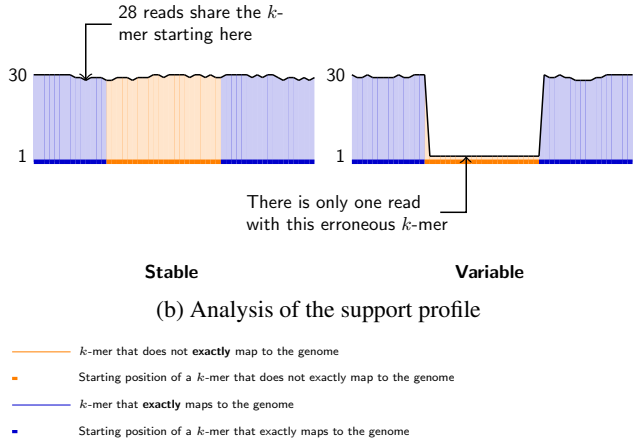
Consider the case of a substitution at position $h$ in the read. All $k$-mers overlapping position $h$ incorporate this difference and will not match the genome. Thus, the location profile will have zero location for $k$-mers starting in the range $[h - k + 1, h]$. Such a range is called a break (Figure 1a). On the contrary, $k$-mers starting left (resp. right) of that range will have one location in the genome region where the RNA comes from. Moreover, locations of the $k$-mers starting in $h - k$ and $h + 1$ are $k + 1$ nucleotides apart on the genome. This allows positioning the difference in both the read and the genome. But that does not help in distinguishing sequencing errors from biological differences (SNP, SNV, editing). The support profile will inform us on this.

If the substitution is a sequencing error, it is with high probability specific to that read. Hence, the $k$-mers overlapping it occur in that read only: their support value is one (minimal). If the substitution is biological, a sizeable fraction of the reads covering this transcript position share the same $k$-mers in that region. Their support remains either similar to that of $k$-mers outside the break or at least quite high depending on the homo-

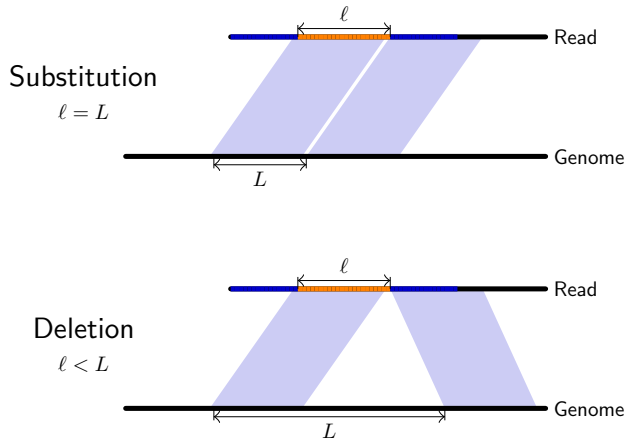
1. Therefore $k$ must be large enough to ensure this. For the human genome, we use $k = 22$.



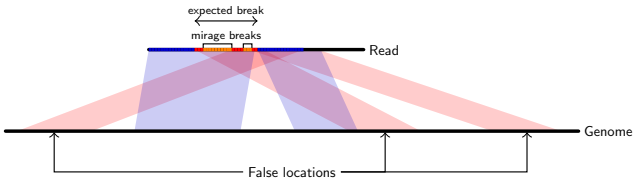

(a) Analysis of the location profile




(b) Analysis of the support profile




(c)




(d)

**Figure 1. CRAC's algorithm.** (a) Illustration of a *break* in the location profile. We consider each $k$-mer of the read and locate it exactly on the genome. In all figures, located $k$-mers are shown in blue, and unmapped $k$-mers in light orange. If the read differs from the genome by, *e.g.* a SNV or an error, then the $k$-mers containing this position are not located exactly on the genome. The interval of positions of unmapped $k$-mers is called a *break*. The end position of the break indicates the error or SNV position. (b) The *support profile*. The support value of a $k$-mer is the number of reads from the collection in which this $k$-mer appears at least once. The two plots show the support profile as a black curve on top of the location profile (in blue and orange). The support remains high (left plot) over the break if many reads covering this region are affected by a biological difference (*e.g.*, a mutation); it drops down in the region of the break when the analyzed read is affected by a sequencing error; in this case, we say the support is *dropping*. (c) Rules for differentiating a substitution, a deletion or an insertion depending on the break. Given the location profile, one can differentiate those by computing the difference between the gap in the genome and the gap in the read between $k$-mers starting before and after the break. (d) False locations and mirage breaks. When false locations occur inside or at the edges of a break they cause *mirage breaks*. False locations are represented in red. The break verification and break merging procedures correct break boundaries (and *e.g.* the correct splice junction boundaries) and avoid detecting a false chimera instead of a deletion.

or hetero-zygosity of the mutation. An artificial difference implies a clear drop in the support profile over the break (Figure 1b). Thus, the ranges of the location break and the support drop will coincide for an error, while a biological difference will not highly alter the support profile over the break. To detect this drop we compare the average support inside versus outside the break using a separation function (Figure 1b). That separation function was determined using an error model observed on Illumina runs. Using this procedure, support profiles are classified as *undetermined* if the support is too low all along the read, and otherwise as either *dropping* or *non-dropping*. Reads with a dropping support profile are assumed to incorporate sequencing errors, and those with a non-dropping support to accurately represent sequenced molecules.

When CRAC processes a read, it first determines $k$-mer locations, analyzes support profile, and apply inference rules whenever possible. Those rules, detailed in the full paper allow one to distinguish substitutions from deletions or insertions (by comparing k-mer positions on the reads and k-mer locations on the genome, see Figure 1c), and to avoid predicting false chimera junctions (Figure 1d). When CRAC is about to predict a chimera (*i.e.* $k$-mer locations are not colinear on the genome, or are too far apart on a same chromosome), the break verification procedure consists of checking that contiguous $k$-mer on the read are also contiguous on the genome (this is equivalent to using a larger $k$). We also check that the break is not too small or cannot be merged with another break, also predicting a chimera, in the read.

The read is classified according to the events (SNV, error, indels, splice, chimera) that are predicted, and its mapping unicity or multiplicity. CRAC algorithm is described for an individual read analysis, but its output can be parsed to count how many reads led to detect the same SNV, indel, splice, or chimera; this can serve to further select candidates. CRAC accepts FASTA/FASTQ formats as input, and outputs distinct files for each class, as well as a SAM formatted file for mapping results.

## 3 Results

In this abstract we focus on results for mapping and predicting splice or chimeric junctions. Please refer to the full publication for more evaluations. Simulated data are compulsory to compute exact sensitivity and accuracy levels, while real data enable us to confront predictions with biologically validated RNAs. For simulating RNA-seq, we first altered a reference genome by random substitutions, indels, and translocations to derive a mutated genome, then reads are "sequenced in silico" by FluxSimulator [3] using annotations and a realistic distribution of expression levels. As read lengths continue increasing, we used two simulated datasets: one (hs75) with a typical read length of 75, another (hs200) with reads of 200 nt representing the future to assess different strategies.

### 3.1 Mapping on current (75 nt) and future (200 nt) reads

Mapping is very common when analyzing reads. Commonly used mappers (Bowtie/BWA/GASSST/SOAP2) compute the best *continuous* alignments up to a certain number of differences [5, 6, 8, 12]. CRAC and GSNAP [15] also consider *discontinuous* alignments to search for reads spanning a splice junction: they can find both continuous and spliced alignments. Note that the former are **not** directly suitable for RNA-seq reads.

Results for mapping 75 nt reads (Table 1a) indicate a high level of precision, but strong differences in sensitivity among tools. Logically, the most sensitive methods are the ones designed for RNA-seq (CRAC and GSNAP), at least 15 points better than their counterparts. That emphasizes the need for specialized mappers.

Analyzing longer reads (200 nt) is another challenge: the probabilities for a read to carry one or several differences (compared to the reference) are higher. In this data set, 36% of the reads cover a splice junction, and 50% carry an error. Compared to 75 nt data, while their precision remains $> 97\%$, all the tools loose in sensitivity (at least 10 points), only CRAC gains 2 points, and remains as precise (Table 1a).

### 3.2 Predicting distinct classes of biological events

Mapping is not a goal *per se*, but only a step in the analysis; the goal of read analysis is to detect candidate biological events of distinct classes (SNV, indels, splice and chimeric junctions) from the reads. The question

| | 75bp | | 200bp | |
|---|---|---|---|---|
| Tool | Sensitivity | Precision | Sensitivity | Precision |
| Bowtie | 75.42 | 99.59 | 55.72 | **99.81** |
| Bowtie2 | 76.64 | 99.26 | 62.31 | 98.78 |
| BWA | 79.29 | 99.13 | 68.66 | 96.86 |
| CRAC | **94.51** | **99.72** | **95.9** | **99.79** |
| GASSST | 70.73 | 99.09 | 59.43 | 97.86 |
| GSNAP | **94.62** | **99.88** | **84.84** | 99.28 |
| SOAP2 | 77.6 | 99.52 | 56.08 | 99.78 |

(a) Comparative evaluation of mapping sensitivity and precision

| | 75bp | | 200bp | |
|---|---|---|---|---|
| Tool | Sensitivity | Precision | Sensitivity | Precision |
| CRAC | 79.43 | **99.5** | **86.02** | **99.18** |
| GSNAP | **84.17** | 97.03 | 72.94 | 97.09 |
| MapSplice | 79.89 | **97.68** | 84.72 | **98.82** |
| TopHat | **84.96** | 89.59 | 54.07 | 94.69 |
| TopHat2 | 82.25 | 92.71 | **88.65** | 91.35 |

(b) Comparative evaluation splice junction prediction tools

| | 75bp | | 200bp | |
|---|---|---|---|---|
| Tool | Sensitivity | Precision | Sensitivity | Precision |
| CRAC | **53.89** | **93.84** | **64.86** | **90.18** |
| MapSplice | 2.33 | 0 | 2.63 | 0.01 |
| TopHat2 | **77.72** | 7.32 | **70.72** | **12.50** |
| TopHat-fusion | 32.73 | 42.02 | | |
| TopHat-fusionPost | 12.26 | **97.22** | | |

(c) Comparative evaluation of chimeric RNA prediction tools

**Table 1. Global mapping sensitivity and precision of 7 tools.** Comparison of **sensitivity** and **precision** of different tools on human simulated RNA-seq (42M-75nt; 48M-200nt) against the Human genome for (1a) mapping, (1b) splice junction prediction, and (1c) chimeric junction prediction. Sensitivity gives the percentage of correctly reported cases over all sequenced cases, while precision gives the percentage of correct cases among all reported cases. Values in bold indicate the maximum of a column. For all tasks on current read length, CRAC combines good sensitivity and very good precision. Importantly, CRAC always improve both figures with longer reads. TopHat-fusion could not process our 200 nt datasets.

is: if this, *e.g.* SNV or splice junction, is present and sequenced, can it be predicted and not buried in a multitude of false positives (FP)?

We assessed splice junction prediction tools, including CRAC, on our simulated datasets. The global comparison and the effect of read length on sensitivity and precision can be observed in Table 1b. With 75 nt, all splice detection tools achieve a good sensitivity, but precision varies by more than 10 points. CRAC reaches 99.5% precision corresponding to 0.5% FP; for comparison, MapSplice and GSNAP output four times as much FP, TopHat 2 about fourteen times and TopHat twenty times more FP. With 200 nt reads CRAC, MapSplice and TopHat 2 improve their sensitivity, while GSNAP and TopHat loose more than 10 points.

Comparison on chimeric RNAs shows that CRAC already offers an acceptable balance between sensitivity and precision with 75 nt reads. TopHat 2 offers a better sensitivity but a much worse precision, while TopHat-fusion and MapSplice sensitivities are much lower (Table 1c). With 200nt reads, CRAC improves sensitivity and loses little in precision, on the other hand TopHat 2 improves precision but loses sensitivity (Table 1c).

### 3.3 Predicting distinct classes of biological events on real data

**3.3.1 Splice junction prediction** To evaluate CRAC's capacity to detect splice junctions in real RNA-seq data, we compared it to state-of-the-art tools (TopHat, GSNAP, MapSplice) on a data set of 75 million stranded 100 nt reads (ERR030856). Splice junctions were searched for using each tool and then confronted to Human

RefSeq transcripts. Each found junction consists of a pair of genomic positions (*i.e.*, the exons 3' end and 5' start) and we consider that it matches a known RefSeq junction if their positions were equal within a 3 nt tolerance. Found junctions were partitioned into *known*, *new* and *other* junctions (KJ, NJ, and OJ respectively). *Known* junctions are those already seen in a RefSeq RNA (*i.e.* junctions between neighboring exons and alternative splicing), *new* ones involve RefSeq exons but in a combination that has not yet been observed in RefSeq, while remaining junctions go into the class *other*. *Novel* junctions will provide new alternative splicing candidates, while junctions in *other* represent totally new candidates RNAs.

MapSplice, GSNAP, CRAC report more than $140,000$ known junctions and all three agree on $126,723$ of them. TopHat reports about $25,000$ junctions less than other tools, and only $1,370$ of its junctions are not detected by any of them. For instance, CRAC covers $93\%$ of TopHat KJ. As known junctions likely contain truly expressed junctions of well studied transcripts, these figures assess the sensitivity of each tool and suggest that in this respect CRAC equals state-of-the-art tools. Logically, the numbers vary more and the agreements are less pronounced among novel junctions. That is why we performed additional tests by BLASTing the junctions on ESTs. Among new and other junctions, BLAST reports good alignments (*i.e.* a low e-value) for respectively $68\%$ and $69\%$ of CRAC's junctions. The corresponding figures are $47\%$ and $47\%$ for GSNAP, $49\%$ and $50\%$ for MapSplice, $51\%$ and $44\%$ for TopHat.

Exploiting simultaneously the genomic locations and support of all $k$-mers gives CRAC some specific abilities for junction detection.

**3.3.2 Comparisons on chimeric splice junction prediction** Edgren *et al.* used deep RNA-sequencing to study chimeric gene fusions in four breast cancer cell lines (BT-474, KPL-4, MCF-7, SK-BR-3), found 3 known cases and validated 24 novel intergenic fusion candidates (*i.e.*, involving two different genes) [1]. As CRAC, TopHat-fusion can predict both intra- and inter-genic chRNA candidates and identify a chimeric junction in a spanning read [4]. For evaluation purposes, we processed each library with TopHat-fusion and CRAC, and compared their results. TopHat-fusion exploits both the read sequence and the read pairs, while CRAC uses only the single read sequence. Otherwise, TopHat-fusion per se and CRAC both select potential chRNAs based on computational criteria. We further filtered out all candidate chimeric reads for which an alternative, colinear alignment was found by GSNAP. Then, filtered predictions were compared and confronted to valid chRNAs. A post-filtering script, called TopHat-fusion-post, based on biological knowledge can be applied to TopHat-fusion results, but in [4] its parameters were chosen "*using the known valid fusions as control*", and may have biased their comparison. So, we remade all predictions using TopHat-fusion with and without TopHat-fusion-post.

The $50$ nt reads, which are well suited for Bowtie-TopHat, represent an unfavorable case for CRAC, which performs better with longer reads. Globally after filtering with GSNAP, TopHat-fusion reports a total of $193,163$ chRNAs, while CRAC outputs $455$: a $600$ fold difference. Therefore CRAC's output has a practicable size and authorizes an in-depth, context dependent investigation to spot promising candidates for validation.

When confronted to the set of validated chimeras of Edgren *et al.* [1], TopHat-fusion and CRAC detect 21 and 20 out of 27, and agrees on 17 of them [2]. Among the 7 cases CRAC misses, only one is a false negative, 4 are unsure (from CRAC's point of view), due to candidates not enough expressed, and no read seems to match the junction of the 2 remaining ones. Considering validated intergenic chRNAs [1], the sensitivity over the 27 valid chRNAs is comparable between TopHat-fusion ($77\% = 21/27$) and CRAC ($74\% = 20/27$), while the precision over the total number of candidates is markedly in favor of CRAC ($21/143003 \simeq 0,01\%$ vs $20/192 \simeq 10,4\%$ [3]).

Notably, among $455$ chRNAs, CRAC reports 36 chRNAs that appear to recur in two, three, or even all four breast cancer libraries. Among these 36 chRNAs: 24 are intra- and 12 are inter-chromosomal, 20 are intragenic, while 16 fuse different genes. Moreover, 35 out of 36 harbor exactly the same junction point in all libraries.

1. TopHat-fusion without the extra post-filtering script.
2. If TopHatFusion-post is applied to TopHat-fusion's results with default parameters, it reports 27 chimera, 11 of them being validated chimeras, which is about half those reported by TopHat-fusion alone.
3. Only intergenic chRNAs are counted here.

Previous investigations [1, 4] did not report any recurrent chRNAs. However, TopHat-fusion also outputs 23 of these chRNAs among 193, 163 candidates.

### 3.4 Time and space consumptions

CRAC requires tens of gigabytes of memory (38 GB for 42 millions reads of length 75, see Table 2) and its running time for analyzing the reads ranges between that of Bowtie and TopHat, which are practical tools. Indexing the Human genome with `crac-index` takes two hours on a x86_64 Linux server on a single thread and uses 4.5 gigabytes of memory.

| Programs | Bowtie | BWA | GASSST | SOAP2 | CRAC | GSNAP | MapSplice | TopHat |
|---|---|---|---|---|---|---|---|---|
| Time (dhm) | 7h | 6h | 5h | 40m | 9h | 2d | 4h | 12h |
| Memory (GB) | 3 | 2 | 43 | 5 | 38 | 5 | 3 | 2 |

**Table 2.** Time and space consumptions for each tool on hs200 dataset, and the human genome. The softwares were launched on a single thread.

## 4 Discussion

CRAC is a multi-purpose tool for analyzing RNA-seq data, that is capable in a single run of predicting sequencing errors, small mutations, normal and chimeric splice junctions (collectively termed events). CRAC is not a pipeline, but a single program that can replace a combination of Bowtie+SAMtools+TopHat/TopHat-fusion, and can be viewed as an effort to simplify HTS analysis. CRAC is neither a mapper, since it uses local coverage information (in the support profile) before computing the genomic position of a read. It implements a novel, integrated approach that draws inferences by analyzing simultaneously both the genomic locations and the support of all $k$-mers along the read. The support of a $k$-mer, defined as the number of reads sharing it, approximates the local read coverage without having the reads mapped. The combined $k$-mers location and support profiles enable CRAC to infer precisely the read and genomic positions of an event, its structure, as well as to distinguish errors from biological events. Integration is not only the key to an accurate classification of reads, but it avoids information loss and saves re-computation, thereby becoming crucial for efficiency. Indeed, CRAC takes more time than state-of-the-art mappers, but is considerably faster than splice junction prediction tools (*e.g.* Bowtie+TopHat). The other key to efficiency is the double indexing strategy: a classical FM-index for the genome and the Gk-arrays for the reads [11]. This makes CRAC's memory requirement higher than that of other tools, but fortunately computers equipped with 64 gigabytes of memory are becoming widespread nowadays. Experiments conducted on simulated data (where all answers are known), which are compulsory to assess a method sensitivity, have shown that CRAC is for each type of prediction at least competitive or surpasses concurrent tools in terms of sensitivity, while it generally achieves better precision. Moreover, CRAC's performances further improve when processing longer reads: *e.g.* on 200 nt reads, 85% sensitivity and 99.3% precision for predicting splice junctions. This is of major importance since Illumina announces a read length of up to 250 nt for its MiSeq [4] and Life Technologies up to 200 nt for its Ion Proton [5].

CRAC analyzes how the location and support profiles vary and concord along the read. Hence $k$-mers serve as seeds (in the genome and in the read set), and $k$ is thus a key parameter. Its choice depends on the genome length [9], and quite conservative values – $k = 22$ for the Human genome – have been used in our experiments. Smaller $k$ are possible with smaller genomes (like bacterial ones). $k$ impacts on the number of false genomic locations (FL) that occurs in the profile; a FL indicates a wrong location for a read $k$-mer, *i.e.* different from the location of origin of the sequenced molecule. This tends to induce an erroneous read mapping or junction border (normal and chimeric junction prediction). However, CRAC uses two criteria to avoid these pitfalls: the *coherence* of locations for adjacent $k$-mers over a range, the *concordance* of locations for the $k$-mers around the break (especially in the break verification and fusion procedures). When $k$-mers surrounding the break have a few, but several, locations, CRAC examines all possible combinations, and as FL occurrences are governed mainly by randomness, this eliminates discordant positions. FL should impact even more the prediction of

4. http://www.illumina.com/systems/miseq/performance_specifications.ilmn
5. http://www.invitrogen.com/1/1/19408-ion-proton-system.html

chimeras. Globally, results on both simulated and real data show that CRAC makes accurate predictions with conservative values. The breast cancer libraries we used have 50 nt reads, but CRAC could still find 74% of the chimeric RNAs validated by Edgren *et al.* [1]. Of course, the $k$ value induces two limitations: first, the minimal exon size detectable in a read is $\geq k$, second, reads must be long enough ($> 40$ nt with $k = 20$ for the Human genome). However, HTS progress towards longer reads, and this should not be a problem.

## Acknowledgments

The authors thank Alban Mancheron for help in packaging CRAC software, Gwenael Piganeau for critical reading of this manuscript. NP is supported by "FONDATION ARC pour la RECHERCHE sur le CANCER" (grant PDF20101202345), "Ligue Contre le Cancer" (grant JG/VP 8102). NP, TC, and ER are supported by a CNRS INS2I (grant PEPS BFC: 66293), the Institute of Computational Biology, Investissement d'Avenir. TC and ER acknoweldge the support from the Region Languedoc Roussillon (grant Chercheur d'Avenir, grant GEPETOS). MS is partially supported by the French ANR-2010-COSI-004 MAPPI Project. TC is supported by "the Canceropole GSO" and the University of Montpellier 2. We acknowledge the funding from Agence Nationale de la Recherche (grant Colib'read ANR-12-BS02-008) and from the MASTODONS Défi from CNRS. All Experiments were run on the ATGC bioinformatics platform http://www.atgc-montpellier.fr/ngs.

## References

[1] Henrik Edgren, Astrid Murumagi, Sara Kangaspeska, Daniel Nicorici, Vesa Hongisto, Kristine Kleivi, Inga H Rye, Sandra Nyberg, Maija Wolf, Anne-Lise Borresen-Dale, and Olli Kallioniemi. Identification of fusion genes in breast cancer by paired-end rna-sequencing. *Genome Biol.*, 12(1):R6, 2011.

[2] P. Ferragina and G. Manzini. Opportunistic data structures with applications. In *Proc. of FOCS*, pages 390–398, 2000.

[3] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic rna-seq experiments with the flux simulator. *Nucleic Acids Res.*, 40(20):10073–10083, 2012.

[4] Daehwan Kim and Steven L Salzberg. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.*, 12(8):R72, 2011.

[5] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

[6] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[7] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[8] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

[9] N. Philippe, A. Boureux, J. Tarhio, L. Bréhélin, T. Commes, and E. Rivals. Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity. *Nucleic Acids Res.*, 37(15):e104, 2009.

[10] N. Philippe, M. Salson, T. Commes, and E. Rivals. CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biology*, 14(3):R30, 2013.

[11] Nicolas Philippe, Mikael Salson, Thierry Lecroq, Martine Leonard, Therese Commes, and Eric Rivals. Querying large read collections in main memory: a versatile data structure. *BMC Bioinf.*, 12(1):242, 2011.

[12] Guillaume Rizk and Dominique Lavenier. GASSST: global alignment short sequence search tool. *Bioinformatics*, 26(20):2534 –2540, 2010.

[13] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[14] K. Wang, D. Singh, Z. Zeng, J.S. Coleman, Y. Huang, G.L. Savich, X. He, P. Mieczkowski, S.A. Grimm, C.M. Perou, J.N. MacLeod, D.Y. Chiang, J.F. Prins, and J. Liu. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, 38(18):e178, 2010.

[15] Thomas D. Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873 –881, 2010.

# New developments in KisSplice: Combining local and global transcriptome assemblers to decipher splicing in RNA-seq data

Alice JULIEN-LAFERRIÈRE[1,2,*], Gustavo AT SACOMOTO[1,2], Rayan CHIKHI[3], Erwan SCAON[3], David PARSONS[2], Marie-France SAGOT[1,2], Pierre PETERLONGO[3], Vincent MIELE[1] and Vincent LACROIX[1,2,*]

[1] Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558; Villeurbanne, France.
[2] BAMBOO, INRIA Grenoble Rhône-Alpes, France
[3] Centre de recherche INRIA Rennes - Bretagne Atlantique, IRISA, Campus universitaire de Beaulieu, Rennes, France
*Corresponding authors: alice.julien-laferriere@univ-lyon1.fr, vincent.lacroix@univ-lyon1.fr

**Abstract** *RNA-seq is deeply changing our way to study transcriptomes. The ultimate goal is to be able to identify and quantify all RNAs present in a sample, even without any prior knowledge of the reference genome, which enables to apply this technology to both model and non model species. However, transcriptome assembly is a difficult task, in particular in the presence of alternative splicing. Two main routes have been followed so far. On the one hand, general purpose transcriptome assemblers [1,2,3] aim at reconstructing all alternative transcripts, but, in order to cope with the inherent combinatorial explosion of this problem, they introduce heuristics which lead them to output only a subset of them (the longest ones). On the other hand, local transcriptome assemblers [4] aim at cataloguing systematically and exactly all the splicing events of a gene but do not provide the full length transcripts. In this work, we propose a pipeline that combines the advantages of both. In practice, we map the output of our local assembler* `KisSplice` *[4] to the output of* `Trinity` *[1] using* `GEM` *[5] and propose a visualisation of the results using* `IGV` *[6]. We also report a major improvement of the memory performances of* `KisSplice` *upon its previous release, thanks to the integration of* `Minia` *[7] for the construction of the de Bruijn graph.*

**Availability** `http://kissplice.prabi.fr`

**Keywords** RNA-seq, alternative splicing, graph algorithms, de-bruijn graph, transcriptome, visualisation

## 1 The model

`KisSplice` is based on the fact that polymorphisms [1] in RNA-seq data create a specific pattern into a *de Bruijn graph* (DBG) associated to the reads of the datasets. For a given value of $k$, a DBG is a directed graph where each node represents a sequence of length $k$, called *$k$-mer*, present in the reads. There is a directed edge between two nodes if the corresponding $k$-mers overlap by $k-1$ nucleotides. The pattern created by polymorphisms in a DBG is called a *bubble*. It corresponds to a pair of (internally) node-disjoint paths with common endpoints.

Fig. 1 shows this non linear structure for an alternative splicing (AS) event. Here the inclusion or exclusion of `S` creates two paths leading from `a` to `b`. The common sites are the two extreme nodes, the shorter path corresponds to junction `ab` whereas the longer path corresponds to the variable part `S` and the junctions `aS`, `Sb`. The shorter path has a predictable length of $2k-2$ nucleotides (concatenation of the $k-1$ nucleotides covering a junction).

1. In this work, we use the term polymorphism when there is an event (at the genomic or transcriptomic level) creating variants in the transcriptome. This covers SNPs, genomic indels and alternative splicing events.

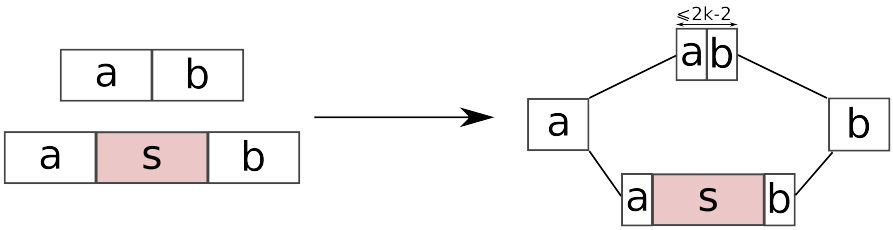



**Figure 1.** Example of a bubble created by an alternative splicing event.

AS bubbles may correspond to exon skipping, intron retention or alternative donor/acceptor site. They however do not cover mutually exclusive exons (the length of the shorter path is larger than $2k-2$) or alternative transcription start/end (the bubble does not close).

Other types of bubbles caused by SNPs, indels and repeats are reported by `KisSplice`, but in the present work we will focus only on AS bubbles, that is, bubbles with a shorter path of at most $2k-2$ nt and a difference of length between the two paths of at least 3nt [2].

## 2 Algorithms at a glance

`KisSplice` identifies and quantifies the polymorphisms in RNA-seq data without a reference genome. A detailed description of the algorithms used in the pipeline are presented in [4]. The pipeline is composed of six main steps: de Bruijn graph construction, biconnected component decomposition, four-nodes compression, bubble enumeration, bubble filtration and classification, read coherence checking and coverage computation.

A polymorphism in the reads creates a non-linear structure in the DBG: a pair of (internally) node-disjoint paths with common endpoints. Note that, when the direction of the edges is disregarded, a bubble corresponds exactly to a simple cycle. Thus in the DBG, ignoring the directions, for any two nodes there is a bubble containing them only if they are in the same *biconnected component* (*BCC*, maximal subgraph such that there is a cycle between any two nodes). Therefore, after the DBG construction, `KisSplice` performs a decomposition into BCCs, with an appropriate graph-traversal algorithm. Since each BCC is completely independent of the others, from now on they are treated in parallel. The next step is four-nodes compression: non-branching bubbles due to SNPs and sequencing errors are detected and compressed, i.e. the two alternative paths are merged into a consensus one. Afterwards, `KisSplice` enumerates the remaining bubbles in each BCC and classifies them into four types of events: alternative splicing (AS), SNPs, small indels and approximate tandem repeats. Finally, the reads are mapped back to each bubble. A bubble is said to be *coherent* if each nucleotide is covered by at least one read. Non-coherent bubbles are discarded, the remaining bubbles are kept and the number of reads mapping to each of them is reported.

Therefore, `KisSplice` uses efficient algorithms to directly output polymorphisms present in RNA-seq data using a DBG structure. Nevertheless it does not reconstruct the full transcriptome, only outputting a context of $2(k-1)$ nucleotides for each bubble.

## 3 Implementation

One of the big challenges of this decade in Bioinformatics is not only to propose algorithms, but also to provide efficient and usable implementations. `KisSplice` follows this line where a specific effort has been made to make it efficient and convenient.

2. To be more precise, AS bubbles also include some indels. In human transcribed regions, 85% of indels concern 1 or 2nt [9]. On the other hand, 99% of AS events are longer than 3nt. If the purpose is really to exclude indels, at the expense of missing some AS events, we recommend to use a threshold of 10nt. Our focus here is to report most AS events. Clearly, between 3 and 10nt, the situation is ambiguous.

### 3.1 Efficient computational footprint

Newest versions of `KisSplice` (from `1.8.0`) integrate improvements to allow any biologist to run it without requiring large computational resources (in terms of time and space). A typical run of `KisSplice` on 100M reads requires only 5GB of RAM.

The de Bruijn graph construction is performed using `Minia` (`http://minia.genouest.org`), which is the up-to-date reference in terms of memory consumption [7,8]. The core code relies on a compact representation of the de Bruijn graph using a Bloom filter, that allows efficient implementation of graph traversal algorithms. `KisSplice` alternates sequential and parallel sections, and future releases will present a significant decrease of the sequential part. `KisSplice` is mainly implemented in C/C++ and the pipeline is driven by a high-level Python "orchestra conductor".

### 3.2 Pipeline combining `KisSplice` with a reference transcriptome: `KisSplice2IGV`

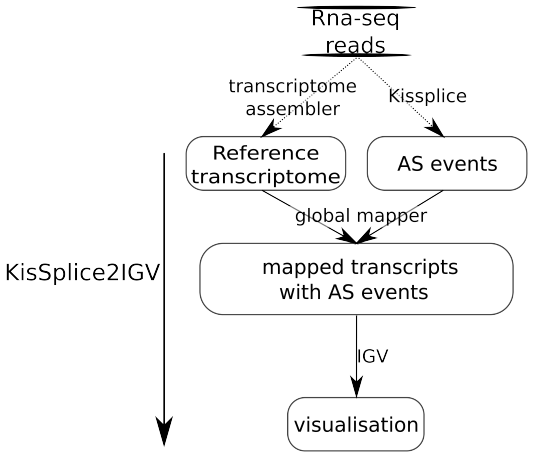



**Figure 2.** Pipeline proposed in the `KisSplice` suite

In this section we propose a new post-treatment of the output of `KisSplice`, combining the sensitivity of `KisSplice` to detect AS events with the longer context given by a full-length transcriptome assembler.

As shown in Fig. 2, we use a transcriptome assembler ( `Trinity`, `Oases`, ...) to infer a reference transcriptome. Alternatively, any reference transcriptome, possibly assembled using other datasets, can be used. The alternative splicing (AS) events reported by `KisSplice` are then mapped to this reference transcriptome using the GEM software [5]. A bubble is considered mapped if at least one path maps to the transcripts. The gapped alignment of the other path can always be deduced from the alignment of the first path. For visualisation, we represent both.

In the case of complex events, involving more than 2 alternative transcripts with a common splice site, `KisSplice` will report all pairs of splice site choices, that is, all pairs of node-disjoint paths in the graph. For instance, if the 4 exons A, B, C and D can be combined in 3 transcripts ABCD, ACD and AD, we will report all pairwise AS events (ABCD Vs ACD, ABCD Vs AD and ACD Vs AD). A path may therefore belong to several bubbles. For instance, the path AD will be present in two bubbles. In the visualisation, we represent it only once. Finally we estimate the coverage of each variant using the read count obtained with `KisSplice` and compute the *reads per kilobase per million mapped read* (*RPKM*), using the following formula $RPKM = RC/((L+r-2k+1)RD)$ , with $L$ the fragment length, $r$ the read length, $k$ the $k$-mer size and $RD$ the read depth. The rationale for the coefficient $(L+r-2k+1)$ comes from the fact that only $L-r+1$ reads can fit in a fragment of length $L$, while $r-1$ reads can overlap this fragment on each side. Since only reads overlapping the fragment with a sufficient length should be accounted for (at least $k$ nt, to ensure that the read covers the variable region of the bubble), $k-1$ are discarded on each side, hence $L-r+1+2(r-1)-2(k-1) = L+r-2k+1$. We represent the alignments using a colour scale depending on the RPKM : the darker the colour is, the more expressed is the gene. An example of the results obtained is shown in Fig. 3.

This post-treatment scheme is implemented as a stand-alone tool `KisSplice2IGV` available on the `KisSplice` website. `KisSplice2IGV` allow direct visualization of the `KisSplice` output along with the results of a transcriptome assembler (or any reference transcriptome obtained independently).

## 4 Calling alternative splicing events in human

We test our pipeline on RNA-seq data from human. Even though we do not require a reference genome to run `KisSplice`, we chose to present the results in the case where one is available because it allows to both discuss the results more in depth and show that `KisSplice` can also be used when a reference genome is available.

We used two sets of reads from the Illumina Body Map 2.0 Project. They consist of 32 M reads from human brain and 39 M reads from liver. Both `KisSplice` and `Trinity` were run with default parameters ($k = 25$). `Trinity` found 52804 components (i.e. genes or gene fragments), out of which 4784 were predicted to have alternative transcripts (both resulting from alternative splicing or alternative transcription). In 3227 cases, `Trinity` outputs only one transcript, while `KisSplice` reported at least one AS event. As we had previously shown in [4], `Trinity` is indeed less sensitive than `KisSplice`and tends to understimate alternative splicing.

On the other hand, there are also some AS events that we know `KisSplice` fails to report. They correspond to AS events in genes flanked by repeats. If the same repeats are shared by many transcribed regions, these AS events are "trapped" in very large biconnected components, for which the enumeration does not terminate. Our current strategy to recover these cases is to increase $k$ and thereby solve more repeats. Clearly, an increase of $k$ leads to a loss a sensitivity. Combining the results obtained for several values of $k$ is therefore a promising strategy, that we did not implement yet.

Fig. 3 is such an example where `Trinity` assembled only one long transcript whereas `KisSplice` detected 3 AS events. Fig. 4 shows the same example, but mapped to the reference genome. Clearly, when `KisSplice` is applied to a non model species, this second visualisation cannot be obtained, but we use it here to explain more in depth the results we obtain.

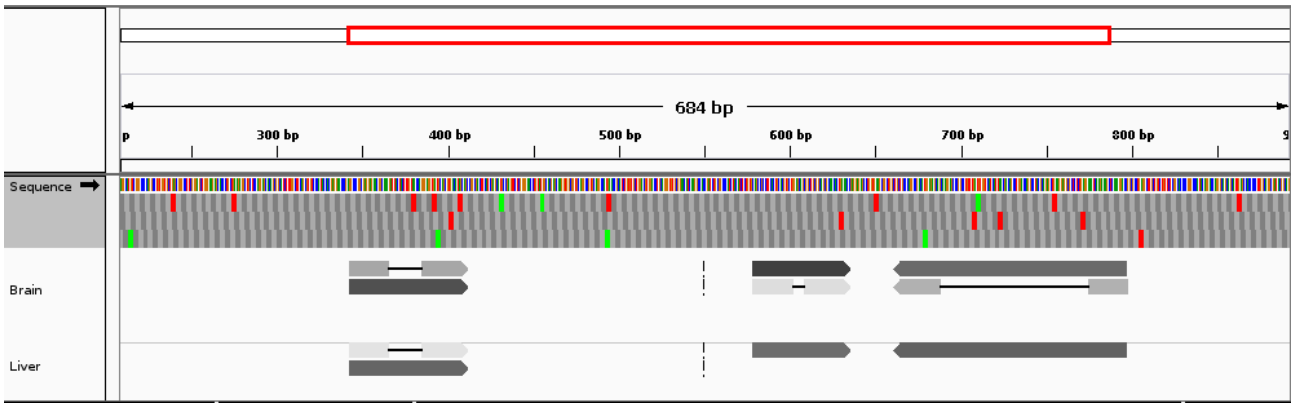



**Figure 3.** Visualisation of the alternative splicing events found by `KisSplice` aligned to one `Trinity` transcript with `IGV`. The first track (named sequence) shows the reference sequence (the `Trinity` transcript) and the three possible open-reading frames (ORF) for the translation. Green squares are initiation codons and red ones are stop codons. Then the two following tracks represent the alignment results, the events reported by `KisSplice`. The colour of an alignment depends on the $\log_{10}$(RPKM). The darker the more expressed. The upper track corresponds to the transcripts found in brain, the lower one to the transcripts found in liver. A thin line indicates a gap in the alignment. Notice that the strand of the `KisSplice` output is not informative.

The new splice variants discovered by `KisSplice` and missed by Trinity clearly correspond to minor isoforms (0.7 to 3 RPKM for novel splice variants versus 21 to 75 for major variant). However, even if they are minor, it is still possible to show that the last one (downstream) is tissue specific (Fisher test, $p$-value $= 0.00239$). The 3 events use canonical splice sites (GT-AG). In the presence of an annotated reference genome, we can further check if these variants are annotated (Fig. 4). In this case, only one out of the 3 events is annotated

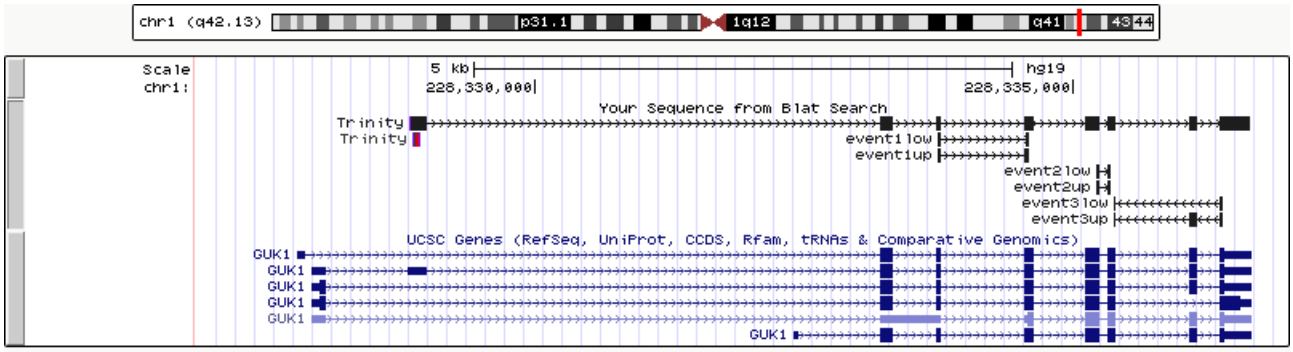



**Figure 4.** Both `Trinity` and `KisSplice` output mapped to the annotated reference genome. Only the third event was annotated.

in UCSC while the 2 others are only confirmed by ESTs (not shown). In the presence of a reference genome, we can also subclassify the events: one of them is an exon skipping event, while the 2 others are alternative acceptor sites.

Furthermore, two of the detected splice variants include variable regions of length not multiple of 3, hence potentially shifting the reading frame. While these events may seem strange at first sight, a more thorough inspection of the ORF (as suggested by the IGV track) reveals that there is an alternative methionine (green) downstream the AS event 1. This methionine could very well be used as an initiation codon, which would then mean that the event falls within a UTR, hence not shifting any reading frame. In this case, this hypothesis seems to be very relevant since the annotations indeed contain an alternative transcript which uses this methionine as an alternative translation start (see Fig. 5). A similar reasoning can be applied to the third event, since the length of the skipped exon is not a multiple of 3. This time, the skipped exon is located at the very end of the ORF, and it does correspond to a shift in the reading frame which ultimately yields a longer protein. These findings actually suggest that finding novel AS events whose length are not multiples of 3 could be used to improve the annotation of alternative translation start and end.

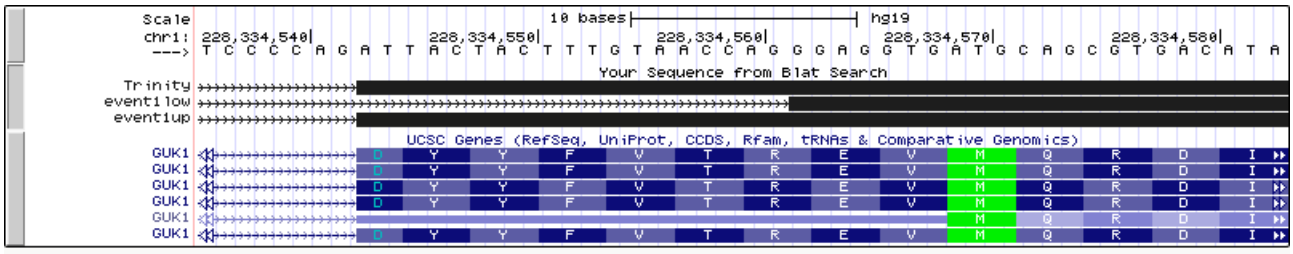



**Figure 5.** Zoom on the first event. In green the possible alternative start codon. The alternative start was already annotated as it can be seen in the 5th transcript.

## 5 Conclusion

We presented a pipeline which enables to analyse the results of `KisSplice` when a reference transcriptome is available. Parts of this pipeline are still under development to be fully automatised but our initial analysis shows that, at least for this gene, events missed by `Trinity` have canonical splice sites, are seen in the annotations or ESTs, and may modify the protein. We had come to similar conclusions in [4] but not with such a level of detail and not with the possibility to visualise the results. We hope that the possibility to visualise the results in a genome browser will help non expert users to easily use `KisSplice` for their RNA-seq analysis. Finally, `KisSplice` can also detect SNPs and indels (there were predicted SNPs for this gene, but we did not include them). In the future, we plan to automatise as well the visualisation of these other types of polymorphism. The joint analysis of both genomic and transcriptomic polymorphisms should prove very useful.

## Acknowledgements

We would like to thank Gunter Roeth (Bull) for interesting discussions about code optimization, Vincent Navratil (Prabi) and Tristan Lefebure (LEHNA) for having deeply tested `KisSplice` on their datasets. This work was funded by the ANR-12-BS02-0008 (Colib'read). The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no [247273]10.

## References

[1] Grabherr MG., Haas BJ., Yassour M., Levin JZ., Thompson DA., Amit I., Adiconis X., Lin F., Raychowdhury R., Zeng Q., and others., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29:644–652, 2011.

[2] Robertson G., Schein J., Chiu R., Corbett R., Field M, Jackman SD., Mungall K, Lee S., Okada HM., Qian JQ., and others., De novo assembly and analysis of RNA-seq data. *Nature methods*, 7:909–912, 2010.

[3] Schulz MH., Zerbino DR., Vingron M., Birney E., Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28:1086–11092, 2012

[4] Sacomoto GAT., Kielbassa J., Chikhi R., Uricaru R., Antoniou P., Sagot MF., Peterlongo P and Lacroix V., KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics*, 13(Suppl 6):S5, 2012.

[5] Marco-Sola S., Sammeth M., Guigo R. and Ribeca P., The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, 9:1185–1188, 2012.

[6] Thorvaldsdottir H., Robinson JT. and Mesirov JP., Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, *Brief. Bioinformatics*, 2012

[7] Chikhi R. and Rizk G., Space-Efficient and Exact de Bruijn Graph Representation Based on a Bloom Filter. *Algorithms in Bioinformatics*, Lecture Notes in Computer Science, 7534:236-248, 2012.

[8] Salikhov K., Sacomoto G., Kucherov G., Using cascading Bloom filters to improve the memory usage for de Brujin graphs. *Arxiv preprint*, arXiv:1302.7278, 2013.

[9] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 2001, 29:308–311.

# Model organisms-oriented metagenomics: recruitment of Illumina reads on marine picocyanobacterial genomes

Gregory FARRANT[1], Erwan CORRE[2], Frédéric PARTENSKY[1], Christophe CARON[2], Silvia ACINAS[3] and Laurence GARCZAREK[1]

[1] Marine Photosynthetic Prokaryotes Team, UMR 7144, Station Biologique, Place Georges Teissier, CS 90074, 29688 Roscoff cedex, France
{gregory.farrant, frederic.partensky, laurence.garczarek}@sb-roscoff.fr
[2] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680 Roscoff, France
{erwan.corre, christophe.caron}@sb-roscoff.fr
[3] Department of Marine Biology and Oceanography, Institut de Ciències del Mar, Consejo Superior de Investigaciones Científicas, Pg Marítim de la Barceloneta 37-49, ES-08003 Barcelona, Spain
sacinas@cmima.csic.es

**Abstract** *As part of the international TARA Oceans project, more than 400 marine metagenomes coming from 152 sampling stations around the world, up to 4 depths and 5 size fractions, have been produced by NGS. Here we present the pipeline that was developed and validated to estimate the diversity, distribution and relative abundance of the marine picocyanobacteria* Synechococcus *and* Prochlorococcus. *This pipeline can be applied to analyze other metagenomes.*

**Keywords** metagenomics, TARA Oceans, *Synechococcus*, *Prochlorococcus*

## Métagénomique d'organismes modèles : recrutement de *reads* Illumina sur des génomes de picocyanobactéries marines

**Résumé** *Dans le cadre du projet international TARA Océans, plus de 400 metagénomes marins issus de 152 stations de prélèvement autour du globe, échantillonnés jusqu'à 4 profondeurs et 5 fractions de taille, ont été produits par NGS. Nous présentons ici le pipeline développé et validé pour estimer la diversité, la distribution et l'abondance relative des clades de picocyanobactéries marines* Synechococcus *et* Prochlorococcus. *Ce pipeline est utilisable pour l'analyse d'autres métagénomes.*

**Mots-clés** metagénomique, TARA Océans, *Synechococcus*, *Prochlorococcus*.

## 1 Introduction

The prokaryotic phylum of Cyanobacteria is the most ancient and likely most diversified group of photosynthetic organisms. The genera *Prochlorococcus* and *Synechococcus* prevail in the euphotic zone of the oceans and play a major role in biogeochemical cycles. Such a ubiquitous distribution implies that they occupy a large variety of niches displaying a wide range of temperatures, quality and quantity of light, nutrient availability, salinity and so on. Such ecosystems diversity is mirrored by a large ecotypic diversity of marine cyanobacteria (Fig. 1).

The vertical distribution of *Prochlorococcus* is known to be tightly linked to light gradients, with 'High-Light' (HL) ecotypes inhabiting the upper mixed layer and 'Low-Light' (LL) ecotypes occupying the base of the euphotic zone [1]. Furthermore, within the HL ecotypes, the HLII clade dominates in permanently stratified waters at low latitudes and is progressively replaced by the HLI clade above 30°N/S [2]. The situation is more complex for the *Synechococcus* genus which presents a much higher level of genetic diversity and for which the relationship between phylogenetic clades and community structure is still poorly understood.

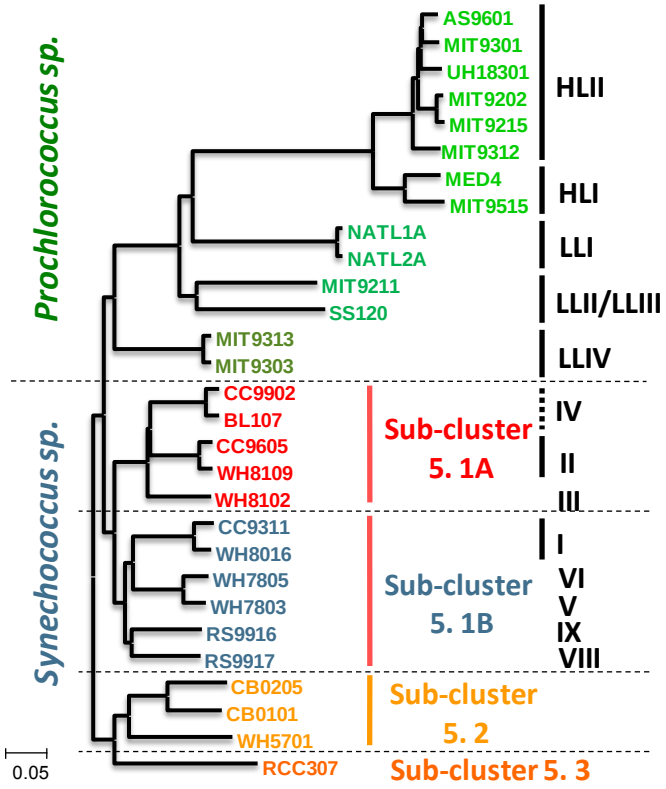



**Figure 1.** 16S rRNA phylogeny of marine picocyanobacteria

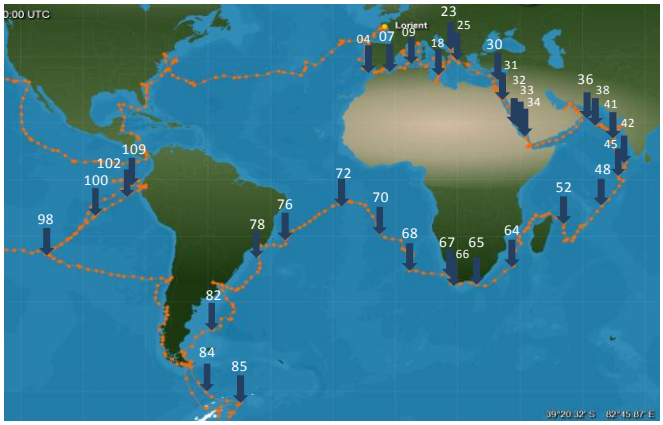


**Figure 2**. The 34 sampling stations of the TARA Oceans program used for this study

In contrast to *Prochlorococcus*, which is mostly influenced by light, *Synechococcus* seems to be more affected by the availability of nutrients, though temperature affects both organisms. Indeed, among the 4 clades dominating *in situ*, clades I, II and IV seem to be more coastal or mesotrophic with clades I and IV dominating in cold water while clade II would correspond to their tropical/subtropical counterpart. In contrast, clade III appears to be confined to the oligotrophic area, but with no temperature preferendum [6]. Still, these parameters are insufficient to fully explain the distribution patterns of the four major ecotypes (I-IV) and we have almost no clue about what parameters control niche partitioning for all other clades.

The TARA Oceans project (Fig. 2, [3]) aims at making a global study of the genetic diversity of marine plankton in order to understand and predict the evolution of the oceanic ecosystem. This 3-year circumnavigating cruise resulted in hundreds of metagenomes that were sequenced at the Genoscope (Evry, France). In the present study, these metagenomes have been analyzed with the aim to better assess the diversity, distribution and relative abundance of *Synechococcus* and *Prochlorococcus* ecotypes. The recruitment pipelines we have developed for that purpose constitute a pertinent cookbook to explore metagenomes looking for a particular set of organisms.

## 2 Materials and Methods

### 2.1 Dataset

A total of 152 stations around the world were sampled at up to 4 depths (surface, deep chlorophyll maximum, oxygen minimum zone and meso-pelagic zone) and up to 5 size fractions (0.2-1.6 µm or 0.2-3 µm, 0.8-5 µm, 5-20 µm, 20-180 µm and 180-2000 µm). Among these, the smallest (bacterial) fraction of 34 stations were sequenced for 2 depths using the Illumina technology, producing hundreds of millions of overlapping pair-end reads. These reads were first assembled using *flash* [4] into ~175 bp reads. Reads were then cleaned, trimmed and/or discarded based on length and quality.

## 2.2 Genome Recruitment to understand distribution and abundance

A total of 57 genomes of picocyanobacteria are currently available (43 *Synechococcus* and 14 *Prochlorococcus*), including 25 yet unpublished *Synechococcus* sequenced at Genoscope. These genomes have been selected to cover as extensively as possible the genetic and ecotypic diversity of these organisms and constitute a representative dataset for genome recruitment.

The first step consists of a *blastn* alignment of all assembled reads against the 57 reference genomes. In order to recruit fragments exhibiting quite low identities with the reference genomes (> 55%) as performed in Ruch et al., 2007, the alignment parameters were set as permissive (-G 8 -E 6 -r 5 -q -4 -W 8). In order to take advantage of all available calculation resources, the sequence sets were atomized using *atomic_blast* (W. Carré, ABiMS, *personal communication*).

Afterwards, two complementary approaches were explored: recruitment plots and profiles of relative abundance of picocyanobacterial clades. The first analysis consists in building graphs representing the best match of all reads on a particular reference genome and the percentage of identity of these alignments [5]. The second analysis needs as a prerequisite to set up a threshold, balanced between specificity and sensibility, above which one can attribute a read to the same ecotype as a given reference genome. Based on recruitment plots, this cut-off was set at 90%. Reads were then tallied up for each genome and eventually summed up for each clade, in order to represent relative clade abundance profiles (Fig. 3).

## 2.3 Marker Recruitment to assess diversity

Genome-based recruitments introduce a bias of representativeness as clades without sequenced representatives cannot be detected. In order to get a better view of the diversity of cyanobacteria in natural samples, reads corresponding to fragments of the *petB* gene, coding for the cytochrome $b_6$, were extracted from the metagenomes. The *petB* gene was selected as a relevant marker of picocyanobacteria diversity, since it is single copy in cyanobacteria, core for photosynthetic organisms, its sequence is short (~540bp) and fairly conserved (allowing an easy automatic ungapped alignment) and however presents a high taxonomic resolution [6]. It is vertically inherited and more than 500 distinct sequences from cultures or environmental samples, including uncultured clades, are available

Two approaches were used to analyze these fragments: i) a targeted approach, where reads were taxonomically attributed to their closest relative based on the latest common ancestor algorithm (LCA) as described by Huson et al. (2007) [7], and ii) a 'blind' approach where sequences were clustered in order to potentially unveil yet uncharacterized clades (Fig. 3).

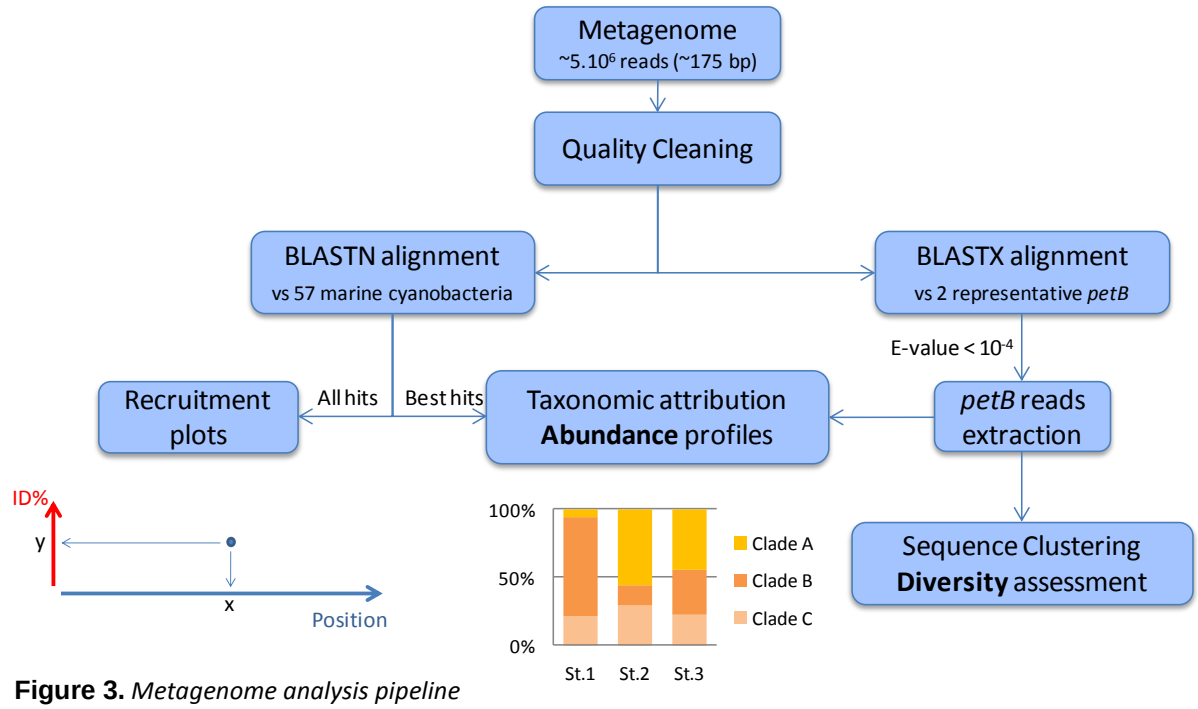



**Figure 3.** *Metagenome analysis pipeline*

## 3 Results and Discussion

Up to 36% of all reads in the bacterial fraction of the metagenomes collected during the TARA-Oceans cruise can be attributed to either *Prochlorococcus* or *Synechococcus*, emphasizing the major role of these organisms in oceanic life. Figure 4 shows that recruitment plots provide good insights about the presence of a given ecotype in a metagenome. Indeed, the *Prochlorococcus marinus* MED4 genome recruits many reads over 90% identity. A gap is observed below 90% showing a break in sequence diversity among marine picoplankton. Two features on the graph deserve particular attention: the region of the chromosome noted A, which recruits many reads with various identity percentages, can be assigned to ribosomal RNAs, whereas the region noted B recruits little or no reads. The latter hypervariable region was previously reported to be an island based on comparative genomics against another *Prochlorococcus* sequence [8]. Genes located in these regions are thought to be involved in adaptation to specific niche in the environment and thus are particularly important to understand the evolution and distribution patterns of marine picocyanobacteria ecotypes. In the example shown in Figure 4, most reads are recruited below 90% identity. This aspecific signal is mainly attributable to other cyanobacteria or more distant relatives.

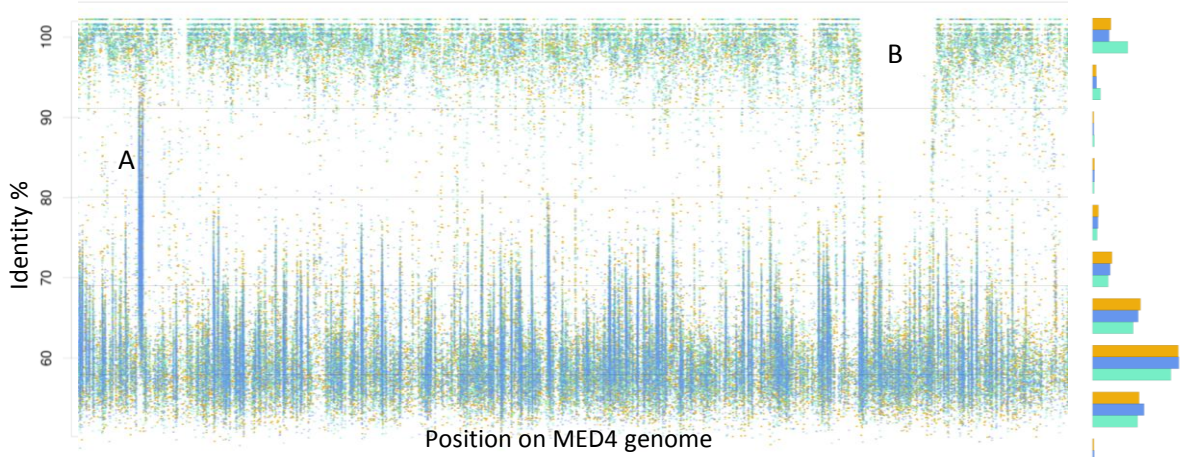



***Figure 4.*** *Recruitment plot of TARA Oceans Stations 7 (turquoise), 23 (blue) and 30 (orange) bacterial fractions against the genome of* Prochlorococcus marinus *MED4. 'A' shows a region recruiting an important amount of reads, and 'B' a region recruiting very few reads. The histogram on the right side shows the abundance of recruited reads along the identity range.*

Profiles of relative clade abundance in TARA Oceans' surface samples are shown in Figure 5. The temperate Mediterranean Sea appears to be dominated by *Prochlorococcus* HLI ecotype, while the subtropical Red Sea and tropical Indian Ocean are dominated by HLII. These results are fully consistent with previous data on the distribution of *Prochlorococcus* ecotypes [2, 9], except for the fairly high relative abundance of LLI ecotypes in surface samples of the Med. Sea and South Atlantic.

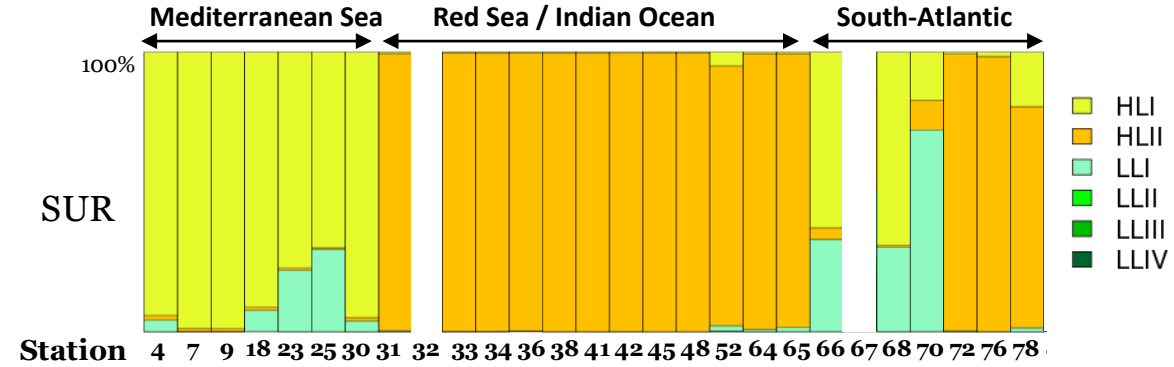



**Figure 5.** Relative abundance of *Prochlorococcus* clades in surface (SUR) samples of TARA Oceans, missing spots correspond to regions of the ocean where the abundance of *Prochlorococcus* is too low to be detected.

Some samples could not be analyzed because of an insufficient abundance of *Prochlorococcus* reads, either because of high latitude or extreme oligotrophy. A cut-off of 100,000 recruited reads, based on average read length and genome size, was chosen to discard these samples. The recruitment of the *petB* marker led to rather similar profiles despite a $10^4$-fold decrease of recruited reads but also allowed to unravel the presence of clades with no representatives in the reference genomes in the TARA data set.

## 4 Conclusions and Perspectives

Genome and genetic marker recruitments constitute complementary approaches to assess the diversity and abundance of the marine picocyanobacteria *Synechococcus* and *Prochlorococcus*. However, some differences between results coming from these two approaches are expected because all genes in a genome do not evolve at the same rate, making the 90% cut-off used in the former approach rather imprecise. Nevertheless the consistence of the profiles obtained using this cut-off with what we know of the distribution of the major clades of *Synechococcus* and *Prochlorococcus* validates this methodology. Pipelines developed in the present study will be used to analyze other size fractions (e.g. to study cyanobacterial symbiosis with eukaryotes) and could also be applied to the study of other metagenomes.

One major challenge that still needs to be addressed is to determine the content of genomic islands in natural populations of picocyanobacteria, as this will provide us with new insights on cyanobacterial adaptation to local niches. Genes present in these metagenomics islands cannot be detected directly by genome recruitment, even when reference genomes used for recruitment originate from the same oceanic region (as shown in Figure 4 where the Mediterranean Sea strain *Prochlorococcus marinus* MED4 was used to recruit reads). However, knowledge of the accessory genomes from the 57 sequenced picocyanobacteria and the fact that the genome context of genomic islands is conserved could be used to specifically retrieve these ‘adaptation genes’ from the metagenomes, though this will require the development of a specific pipeline.

## Acknowledgements

This work is supported by the ANR Pelican ANR-09-GENM-030) and the European Union's Seventh Framework Programs FP7 MicroB3 and MaCumba (grant agreements 287589 and 311975, respectively). We warmly thank members of the TARA Oceans consortium as well as the Genoscope.

## References

[1] Moore LR, Rocap G and Chisholm SW. *Physiology and molecular phylogeny of coexisting* Prochlorococcus *ecotypes*. Nature, Lond 393:464-467, 1998.

[2] Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM and Chisholm SW Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* 311:1737-1740, 2006.

[3] Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, et al. *A holistic approach to marine eco-systems biology*. PLoS Biol 9: e1001177, 2011.

[4] Magoč T., Salzberg S. L., *FLASH*: *Fast length adjustment of short reads to improve genome assemblies.* Bioinformatics 27, 2957, 2011.

[5] Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. *The* Sorcerer II *Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific*. PLoS Biol 5(3): e77., 2007.

[6] Mazard S, Ostrowski M, Partensky F, Scanlan DJ., *Multi-locus sequence analysis, taxonomic resolution and biogeography of marine Synechococcus.* Environ Microbiology; 14(2):372-86, 2012.

[7] Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). *MEGAN analysis of metagenomic data.* Genome Res 17, 377-386

[8] Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF and Chisholm SW *Genomic islands and the ecology and evolution of* Prochlorococcus. Science 311:1768-1770, 2006.

[9] Mella-Flores D, Mazard S, Humily S, Partensky F, Mahé F, Bariat L, Courties C, Marie D, Ras J, Mauriac R, Ostrowski M, Scanlan DJ and Garczarek L *Is the distribution of* Prochlorococcus *and* Synechococcus *ecotypes in the Mediterranean Sea affected by global warming?* Biogeosciences 8:2785-2804, 2011.

# Assessing the enrichment significance of a Position Weight Matrix (PWM) along a DNA sequence

Julien DUMAZERT[1], Jean-Yves STEPHAN[1], Marie-Agnès PETIT[2] and Sophie SCHBATH[3]

[1] Ecole Polytechnique, Palaiseau, France
{julien.dumazert, jean-yves.stephan}@polytechnique.edu
[2] Micalis, UMR1319 INRA, Jouy-en-Josas 78352 cedex, France
{marie-agnes.petit}@jouy.inra.fr
[3] Mathématique, Informatique et Génome, UR1077 INRA, Jouy-en-Josas 78352 cedex, France
{sophie.schbath}@jouy.inra.fr

**Abstract** *We present a novel statistical approach to evaluate if a given Position Weight Matrix (PWM) is significantly enriched in a given sequence. We define the weighted count of a PWM in a sequence without choosing any arbitrary threshold and we propose a compound Poisson approximation for the weighted count distribution, which appears more accurate than a Gaussian distribution. Our method, called PWMstat, is based on an efficient algorithm to simulate the ad-hoc compound Poisson distribution and provides then an enrichment $p$-value. By comparison with Bioprospector, an existing motif discovery tool, we obtained that Bioprospector scores do not generally reflect well the enrichment significance of a PWM. Our method is illustrated on the Noc binding site in Bacillus subtilis.*

**Keywords** Position Weight Matrix, PWM, DNA motifs, count statistics, overlapping occurrences, $p$-value, compound Poisson distribution.

## 1 Introduction

Position weight matrices (PWMs) are commonly used to represent DNA motifs and more particularly transcription factor binding sites. They allow to capture important degeneracy of DNA motifs and then to describe binding preferences more precisely than words. A PWM of length $\ell$ is indeed a $4 \times \ell$ matrix that contains the probabilities of the four bases for each of the $\ell$ positions of the motif. PWMs can be built from motif alignments or generated by motif discovery algorithms (eg. MEME ([2]), Bioprospector ([4]) or XXmotif ([3])) which look for significant PWM enrichment.

Few statistical methods exist to compute the significance of the enrichment of a given PWM in a given sequence. Indeed, classical methods designed for word counts ([7], [10]) are not efficient because they require the enumeration of all words compatible with the PWMs and, usually, this set is too large. Dedicated methods are then necessary to directly handle PWMs. Most of existing methods focusing on PWMs enrichment define an occurrence (or a hit) of a PWM in a sequence as a sequence position whose affinity score with the PWM is higher than a certain threshold $\alpha$, and aim at studying the occurrence probability ([12]) or the distribution of the number of occurrences ([6]). Despite the problem of choosing the threshold $\alpha$, previous methods consider all occurrences as similar for the enrichment problem, independently of the value of the affinity scores given by the PWM.

We propose a novel approach which generalizes the definition of a word count, namely a sum over all sequence positions of binary variables equal to 1 if the word occurs at that position or 0 otherwise. Indeed, we sum the affinity scores of the PWM at each position in the sequence, defining then a weighted count of a PWM in a given sequence. Compatible words with a low affinity score will then contribute less than the words with a higher affinity score. If the PWM is reduced to a unique compatible word, then the weighted count of the PWM is equal to the compatible word count; our approach is then a natural generalization of word count statistics.

Note that [5] and [9] also propose a threshold-free approach to model the affinity of transcription factors with a given sequence region; their strategy is very different from ours because, instead of looking for occur-

rences of transcription factors, they rather seek an appropriate normalisation which allows them to compare the affinities of different transcription factors directly with each other.

In the next section we will mathematically define the weighted count $T$ of a PWM in a sequence and Section 3 will be devoted to the approximation of the probability distribution of $T$, which will provide a $p$-value assessing if the sequence is significantly enriched in the PWM. We first compute the exact expectation and variance of $T$ under a random Bernoulli sequence and show a Gaussian approximation (Section 3.1). However, for PWMs having lots of compatible words with very small affinity scores, the $T$ statistics is better approximated by a compound Poisson variable (Section 3.2). We developped an efficient method to compute the approximating compound Poisson distribution and the $p$-value associated to the observed weighted count of the PWM (Section 4). We then illustrate our method, called PWMstat, on the Noc protein binding site in *Bacillus subtilis* genome (Section 5.1) and show that our enrichment $p$-value is not equivalent to the score provided by Bioprospector to rank its outputed PWMs (Section 5.2). It means that motifs found with Bioprospector are not necessarily ranked according to the significance of their weighted count in the sequence.

## 2 Weighted Count of a PWM in a Given Sequence

Let $\mathbf{m} = (m_{a,j})_{(a,j)\in\{a,c,g,t\}\times\{1,\dots,\ell\}}$ denote a position weight matrix of length $\ell$ on the 4-letter DNA alphabet $\mathcal{A} = \{a, c, g, t\}$. Such a matrix $\mathbf{m}$ defines a particular set $\mathcal{W}$ of weighted $\ell$-letter words, where the weight of a $\ell$-letter word $\mathbf{w} = w_1 w_2 \cdots w_\ell$ is given by $\nu(\mathbf{w}) := \prod_{j=1}^{\ell} m_{w_j,j}$. $\mathcal{W}$ is then the set of all the $\ell$-letter words whose weight is stricly positive; words from $\mathcal{W}$ are then said "compatible" with the matrix $\mathbf{m}$.

*Affinity score* Consider a sequence $S = (S_1, S_2, \dots, S_n)$ of $n$ letters in the alphabet $\mathcal{A}$. We define the affinity score $\nu_i$ between the matrix $\mathbf{m}$ and the sequence at position $i$ as the weight of $S_i S_{i+1} \cdots S_{i+\ell-1}$, namely:

$$\nu_i = \prod_{j=1}^{\ell} m_{S_{i+j-1},j}. \tag{1}$$

*Weighted count* Now, to measure how frequent is a position weight matrix $\mathbf{m}$ along a sequence $S$, we will study the following statistics:

$$T = \sum_{i=1}^{n-\ell+1} \nu_i. \tag{2}$$

This statistics has the advantage of taking all the occurrences of all compatible words into account and to weight them according to the matrix $\mathbf{m}$. It is therefore different from previous approaches consisting in counting the number of occurrences of compatible words whose weigth exceeds a given threshold ([6]). Both approaches are a generalization of the classic word count in case the matrix reduces to a unique compatible word (with weight 1).

The question addressed in this paper is "How to assess the statistical significance of the observed value $t$ of the weighted count $T$ ?", or in other words "What is the statistical distribution of $T$ in a random sequence ?" so that we could compute the $p$-value $\mathbb{P}(T > t)$.

## 3 Distribution of the Weighted Count

Because of the way the position weigth matrix is constructed – independence of the positions – we will consider Bernoulli sequences: independent letters with probabilities $(p(a), p(c), p(g), p(t))$ fitted to the DNA sequence composition of interest. However, the method can be generalized to Markovian sequences.

A naive approach for computing $\mathbb{P}(T > t)$ could consist in merely simulating a large number of Bernoulli sequences and determining the empirical distribution of $T$, but it would require too mmany simulations to accurately identify very exceptional events ($p$-values very close to 0). This naive approach will however be performed in our analysis only to measure the quality of the two approximations we propose instead: a Gaussian

approximation (Section 3.1) and a compound Poisson approximation (Section 3.2). Theses two asymptotic regimes are well known in the theory of counting statistics for clumping events ([7], [8]) and roughly correspond to expectedly frequent versus rare motifs.

### 3.1 Gaussian Approximation

The asymptotic normality of the weighted count $T$ follows from a straightforward application of the Central Limit Theorem for ergodic Markov chains. Indeed, $T$ is a sum of random variables $\nu_i$ (see equation (2)) which can be written as a certain function $f$ of $(S_i, S_{i+1}, \ldots, S_{i+\ell-1})$; since the sequence letters are independent, $(\nu_i)_i$ forms an homogeneous and ergodic Markov chain of order $\ell - 1$.

We could then approximate the distribution of $T$ by the Gaussian distribution with expectation $\mathbb{E}(T)$ and variance $\text{Var}(T)$, where the formulas for expectation and variance of $T$ are given below.

*Expectation of $T$* From equation (2), we have that $\mathbb{E}(T) = (n - \ell + 1)\mathbb{E}(\nu_1)$ and from the affinity score definition (1), we have

$$\mathbb{E}(\nu_1) = \prod_{i=1}^{\ell} \big(p(a)m_{a,i} + p(c)m_{c,i} + p(g)m_{g,i} + p(t)m_{t,i}\big).$$

*Variance of $T$*

$$\text{Var}(T) \quad = \quad \sum_{i=1}^{n-\ell+1} Var(\nu_i) + \sum_{i \neq j \in \{1,\ldots,n-\ell+1\}} Cov(\nu_i, \nu_j)$$

Since the variables $(\nu_i)_i$ are identically distributed and $\nu_i$ is independent of $\nu_j$ when $|i - j| \geq \ell$, we get:

$$\begin{aligned}
\text{Var}(T) \quad &= \quad (n - \ell + 1)\text{Var}(\nu_1) + 2 \sum_{i=1}^{n-2\ell+2} \sum_{k=1}^{\ell-1} \text{Cov}(\nu_i, \nu_{i+k}) + 2 \sum_{i=n-2\ell+3}^{n-\ell+1} \sum_{k=1}^{n-i-\ell+1} \text{Cov}(\nu_i, \nu_{i+k}) \\
&= \quad (n - \ell + 1)\mathbb{E}(\nu_1^2) + 2(n - 2\ell + 2) \sum_{k=1}^{\ell-1} \mathbb{E}(\nu_1 \nu_{1+k}) + 2 \sum_{i=n-2\ell+3}^{n-\ell+1} \sum_{k=1}^{n-i-\ell+1} \mathbb{E}(\nu_i \nu_{i+k}) \\
&\qquad - \Big( (n - \ell + 1) + (\ell - 1)(2n - 3\ell + 2) \Big) \mathbb{E}(\nu_1)^2.
\end{aligned}$$

And we have a following straightforward formula for joint expectations for $j \in 0, \ldots, \ell - 1$

$$\begin{aligned}
\mathbb{E}\left(\nu_1 \nu_{1+j}\right) \quad &= \quad \prod_{i=1}^{j} (p(a)m_{a,i} + p(c)m_{c,i} + p(g)m_{g,i} + p(t)m_{t,i}) \\
&\times \quad \prod_{i=j+1}^{\ell} (p(a)m_{a,i}m_{a,i-j} + p(c)m_{c,i}m_{c,i-j} + p(g)m_{g,i}m_{g,i-j} + p(t)m_{t,i}m_{t,i-j}) \\
&\times \quad \prod_{i=\ell-j+1}^{\ell} (p(a)m_{a,i} + p(c)m_{c,i} + p(g)m_{g,i} + p(t)m_{t,i})
\end{aligned}$$

Computation of the parameters of the Gaussian approximation is therefore very easy with those formulas and we have access to the *p*-value under the Gaussian approximation as follows:

$$\mathbb{P}(T > t) \simeq 1 - \Phi\left(\frac{t - \mathbb{E}(T)}{\sqrt{\text{Var}(T)}}\right) \tag{3}$$

where $\Phi$ is the cumulative distribution function of $\mathcal{N}(0, 1)$.

### 3.2 Compound Poisson Approximation

Since occurrences of compatible words tend to occur in clumps of overlapping occurrences, compound Poisson distribution is generally relevant to model such clumping effect ([11], [10]).

A clump is a series of overlapping compatible words. We give a special importance to the first word of the clump: a clump beginning with the word $\mathbf{w}$ is called a clump of type $\mathbf{w}$. A clump can be made of a single word, if that word is not overlapped by another compatible word.

By gathering the occurrences of all the compatible words in the sequence into clumps, we can rewrite our statistic $T$ like:

$$T = \sum_{i=1}^{n-\ell+1} \nu_i = \sum_{w \in \mathcal{W}} \sum_{c=1}^{C_{\mathbf{w}}} X_c^w \tag{4}$$

where $C_{\mathbf{w}}$ is the number of clumps of type $\mathbf{w}$ and $X_c^{\mathbf{w}}$ is the sum of the affinity scores of the words of the $c$-th clump of type $\mathbf{w}$. The real advantage of this decomposition is that the random variables $(X_c^{\mathbf{w}})_c$ are now independent (clumps are disjoint and the sequence letters are independent) and identically distributed for a given $\mathbf{w}$. Contrarily to the word case ([11], [10]), the $X^{\mathbf{w}}$ distribution cannot be reached analytically but the important point of this paper is that we have designed an algorithm to efficiently sample the distribution of the variables $X^{\mathbf{w}}$ (see Section 4). As for the word case, we use combinatorics of overlaps between compatible words.

Moreover, we have shown by using the Chen-Stein method ([1]) that the number $C_{\mathbf{w}}$ of clumps of type $\mathbf{w}$ can be approximated by a Poisson variable with expectation $\mathbb{E}(C_{\mathbf{w}})$ (not shown in this abstract) and we checked this theoretical approximation thanks to naive simulations. The expected number of clumps of type $\mathbf{w}$ is given below.

$$\begin{aligned}
\mathbb{E}(C_{\mathbf{w}}) &= (n-\ell+1) \times \mathbb{P}(\text{a clump of type } \mathbf{w} \text{ starts in position } i) \\
&= (n-\ell+1) \times \mathbb{P}(\text{a word } \mathbf{w} \text{ starts in position } i) \times p^{\mathbf{w}} \\
&= (n-\ell+1) \times \prod_{j=1}^{\ell} p(w_j) \times p^{\mathbf{w}}
\end{aligned}$$

where

$$\begin{aligned}
p^{\mathbf{w}} &= \mathbb{P}(\text{No compatible word starts in position } i-\ell+1, ..., i-1 \mid \mathbf{w} \text{ starts in position } i) \\
&= 1 - \sum_{\mathbf{w}' \in \mathcal{W}} \frac{p(\mathbf{w}')}{p(\mathbf{w})} p_{\mathbf{w}'\mathbf{w}}
\end{aligned}$$

and the probability $p_{\mathbf{w}'\mathbf{w}}$, which represents the probability that $\mathbf{w}$ is the nearest word overlapping $\mathbf{w}'$ on the right (see Section 4.1 for a formal definition), can be calculated thanks to the algorithm proposed in Section 4.1.

Finally, we can get an approximate empirical distribution for $T$, and then an approximate $p$-value for the observed weighted count $t$, by sampling, for each compatible words $\mathbf{w}$, a number of clumps $C_{\mathbf{w}}$ from the corresponding Poisson distribution and then, for each clump, from the distribution of $X^{\mathbf{w}}$. This way of getting the empirical distribution of $T$ is intermediate between the naive simulation approach, consisting of simulating random sequences and computing $T$ for each sequence, and an analytical approach as we could have for word counts. The great practical advantage of our compound Poisson approximation, called PWMstat, is that it is much faster than the naive simulation, as it will be illustrated in the Application Section.

## 4 Getting the Exact Distribution of the Affinity Score $X^{\mathrm{w}}$ of Clumps of Type $\mathrm{w}$

The first algorithm that we describe exhaustively explore all the compatible words and their possible overlaps.

### 4.1 Pre-treatment: Probability of Overlaps

In this section, we give an algorithm to compute the probability that an occurrence of a given compatible word is overlapped by another occurrence of a compatible word on the right. This probability will be needed to compute the distribution of $X^{\mathbf{w}}$ and is required to compute the parameter of the Poisson distribution modelling the number of clumps of a given type (see Section 3.2). The calculation of this probability consists in an exhaustive exploration of all the compatible words and their possible overlaps.

Let $\mathbf{w}$ and $\mathbf{w}'$ be two compatible words of $\mathcal{W}$. We say that $\mathbf{w}'$ is the *nearest overlapping word on the right* of $\mathbf{w}$ if word $\mathbf{w}'$ starts before the end of word $\mathbf{w}$ and no other word starts between $\mathbf{w}$ and $\mathbf{w}'$. More formally, assume that word $\mathbf{w}$ starts at position $i$; $\mathbf{w}'$ is the *nearest overlapping word on the right* of $\mathbf{w}$ if there exists $k \in 1, \ldots, \ell - 1$ so that

$$s_{i+k}s_{i+k+1}\ldots s_{i+k+\ell-1} = \mathbf{w}' \text{ and } \forall j,\ i < j < i + k,\ \nu_j = 0.$$

The probability of this event is denoted by $p_{\mathbf{ww}'}$. Note that it does not depend on the position $i$ at which word $\mathbf{w}$ starts since we use an homogeneous Bernoulli model for the letter sequence.

EXAMPLE 4.1. *Consider the following PWM that will be used throughout the paper to illustrate concepts and notions:*

$$\begin{pmatrix} 1 & 0.5 & 0.5 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0.5 & 0.5 \end{pmatrix} \tag{5}$$

*We have $\mathcal{W} = \{AAA, AAT, ATA, ATT\}$. Below is an example of 3-word clump of type AAT. A clump can be described by its first word and the sequence of nearest overlapping words on the right. Here: AAT, ATA,ATT.*

...GTAATATTACG...

The computation of $p_{\mathbf{ww}'}$ is described by Algorithm 1. For each possible shift between 1 and $\ell - 1$, it checks whether it is possible to create compatible words starting with the end of word $\mathbf{w}$. Our implementation is recursive and only enumerates compatible words (a brute force algorithm would create all possible sequences of length $2\ell-1$ and select those presenting overlapping words). In the pseudocode of algorithm 1, the recursion is hidden in the **for all** loops.

**Algorithm 1** Compute $p_{\mathbf{ww}'}$

**Ensure:** $\forall \mathbf{w}, \mathbf{w}' \in \mathcal{W},\ \text{overlaps}[\mathbf{w}][\mathbf{w}'] = p_{\mathbf{ww}'}$
  **for all** $\mathbf{w} \in \mathcal{W}$ **do**
    **for all** $\mathbf{w}' \in \mathcal{W}$ **do**
      overlaps[$\mathbf{w}$][$\mathbf{w}'$] $\Leftarrow 0$
    **end for**
    **for** $k = 1 \rightarrow \ell - 1$ **do**
      **if** $\Pi_{i=1}^{\ell-k} m_{w_{k+i},i} \neq 0$ **then**
        **for all** $s_1 \ldots s_k$ **so that** $\mathbf{w}' = w_{k+1}\ldots w_\ell s_1 \ldots s_k \in \mathcal{W}$ **do**
          **if** $\forall j,\ 1 < j < k + 1,\ \nu_j = 0$ **then**
            overlaps[$\mathbf{w}$][$\mathbf{w}'$] $\Leftarrow$ overlaps[$\mathbf{w}$][$\mathbf{w}'$] $+ \Pi_{i=1}^{k} p(s_i)$
          **end if**
        **end for**
      **end if**
    **end for**
  **end for**

### 4.2 From Clump Trees to Clump Distribution

A clump can be described by the sequence of the words composing it (see Example). Except the first one, each word is the *nearest overlapping word on the right* of the preceding word. Therefore, probabilities $p_{\mathbf{ww}'}$

defined in Section 4.1 suffice to perform calculations of expectation and simulations of affinity score $X^{\mathbf{w}}$ of clumps of type $\mathbf{w}$.

For instance, in order to find the probability that a clump of type $\mathbf{w}_1$ is the finite sequence of compatible words $(\mathbf{w}_1, \ldots, \mathbf{w}_n)$, we write

$$\begin{aligned}\mathbb{P}\left(\mathbf{w}_2, \ldots, \mathbf{w}_n | \text{ clump of type } \mathbf{w}_1\right) &= \left(\prod_{i=2}^{n} \mathbb{P}\left(\mathbf{w}_i \text{ follows } \mathbf{w}_{i-1}\right)\right) \mathbb{P}\left(\text{ no compatible word follows } \mathbf{w}_n\right) \\ &= \left(\prod_{i=2}^{n} p_{\mathbf{w}_{i-1}\mathbf{w}_i}\right)\left\{1 - \sum_{\mathbf{w}\in\mathcal{W}} p_{\mathbf{w}_n\mathbf{w}}\right\}\end{aligned}$$

Using the representation of clumps as sequences of *nearest overlapping words on the right*, the set of clumps of a given type is naturally endowed with a tree structure. The nodes of the tree are the compatible words of set $\mathcal{W}$, and the branch between a parent $\mathbf{w}$ and a child $\mathbf{w}'$ is weighted with probability $p_{\mathbf{w}\mathbf{w}'}$. A finite path in the tree is a possible clump and using simple rules, it is possible to compute its probability of appearance.

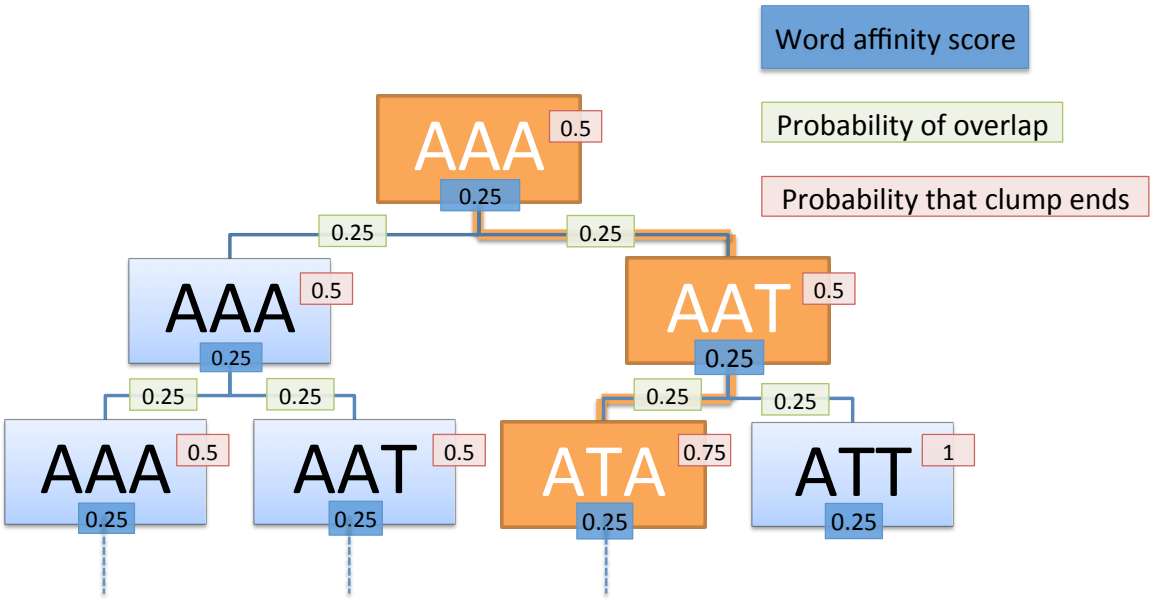



**Figure 1.** Clump tree for PWM given in equation (5)

Fig. 1 shows the tree of clumps of type AAA for the PWM given in Example 1 (see (5)) and equiprobable letters in the genome. It gives an example of a possible 3-word clump (AAATA). It is represented by a finite path in the tree (orange). Its affinity score is the sum of the scores of the three word composing it ($0.25 + 0.25 + 0.25 = 0.75$) and its probability among the set of AAA-type clumps is the product of the weights of its branches and of the probability that the clump ends on its last word ATA ($0.25 \times 0.25 \times 0.75 = 3/64$).

It is possible to derive analytic formulas for affinity score of clumps from those tree representations. The idea behind the method is that clump trees contain redundant information (repeated subtrees). Nevertheless, the computational requirements of analytic resolution are too high for real-life PWMs and we opted for a very simple simulation procedure instead.

To simulate the affinity score $X^{\mathbf{w}}$ of a clump of type $\mathbf{w}$, our algorithm follows a path in the clump tree associated to $\mathbf{w}$. Starting from the root, it chooses at each level to follow a branch –according to its probability– or to end the clump at this level. The simulation process is formally described by algorithm 2 where the set of compatible words $\mathcal{W}$ is identified to $\{1, ..., |\mathcal{W}|\}$ in order to simplify notation.

**Algorithm 2** Returns a possible value of $X^{\mathbf{w}}$

**Require:** a word $\mathbf{w}$
**Ensure:** $x$ contains a possible value of $X^{\mathbf{w}}$
  $x \Leftarrow 0$
  $\mathbf{w}_1 \Leftarrow \mathbf{w}$
  continue $\Leftarrow true$
  **while** continue **do**
    $x \Leftarrow x + \nu(\mathbf{w}_1)$
    $p \Leftarrow$ random number between 0 and 1
    **if** $p > \sum_{\mathbf{w}' \in \mathcal{W}} p_{\mathbf{w}_1 \mathbf{w}'}$ **then**
      continue $\Leftarrow false$
    **else**
      $\mathbf{w}_1 \Leftarrow \mathbf{w}'$ such as $\sum_{v=1}^{\mathbf{w}'-1} p_{\mathbf{w}_1 v} < p \leq \sum_{v=1}^{\mathbf{w}'} p_{\mathbf{w}_1 v}$
    **end if**
  **end while**

## 5 Application

We will first apply our approach to evaluate the significance of the Noc protein binding site enrichment in *Bacillus subtilis* genome; Moreover, we will compare, on this example, the quality of the compound Poisson approximation and the Gaussian approximation thanks to the empirical distribution of the weighted count $T$ obtained via naive simulation (section 5.1).

In section 5.2, we will compare the PWMstat $p$-value of the 20 PWMs with highest Bioprospector score on a given sequence.

### 5.1 Noc Binding Site in *B. subtilis*

Rod shaped bacteria use a set of proteins collectively designated as the divisome to initiate cell division. This division occurs at mid-cell of the elongated mother cell, and genetic experiments have shown that two systems ensure that the divisome forms precisely at mid-cell and nowhere else. One of these systems is based on the Noc protein in *B. subtilis*, which prevents divisome assembly at the vicinity of the chromosome ([13]). The Noc barrier is simply due to the fixation of Noc proteins all along the two replicating sister chromosomes, and ChIP on chip experiments allowed to determine that the fixation of Noc to the chromosome is mediated by a 14 bp-long, palindromic motif (see Fig. 2), called NBS, which is present 74 times along the chromosome, and absent from the terminus region ([14]). The arrival of the Noc-free terminus region at mid-cell for its replication may therefore function as a signal to start divisome assembly.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 88.43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 88.43 |
| C | 0 | 0 | 0 | 11.41 | 100 | 87.77 | 73.37 | 73.37 | 87.77 | 100 | 11.41 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 11.41 | 0 | 0 | 11.41 | 11.41 | 0 | 0 | 11.41 | 0 | 0 | 0 |
| T | 11.57 | 100 | 100 | 77.18 | 0 | 12.23 | 15.22 | 15.22 | 12.23 | 0 | 77.18 | 100 | 100 | 11.57 |

**Figure 2.** Position weight matrix representing the 14 bp-long NBS motif; probabilities are expressed in percentage.

We then considered the DNA sequence composed of the concatenated ChIP regions from *B. subtilis* genome (569806 bps) and computed the weighted count $T$ of the NBS motif: we got $T = 0.0420475$. The $p$-value obtained with our compound Poisson approximation ($10^6$ sampling from the $X$ distribution) is equal to 0.003812 and its 95% confidence interval is $[0.003691; 0.003933]$. As we can see on Fig. 3, the quality of the compound Poisson aproximation is very good as it is very close to the empirical distribution. The empirical distribution has been computed thanks to the simulation of 20,000 Bernoulli sequences with same composition than our DNA sequence of interest (it took 47 minutes which is definitely too much compared to the 107 seconds required for the $10^6$ samplings of PWMstat). The Gaussian distribution is particularly not adapted to this PWM.

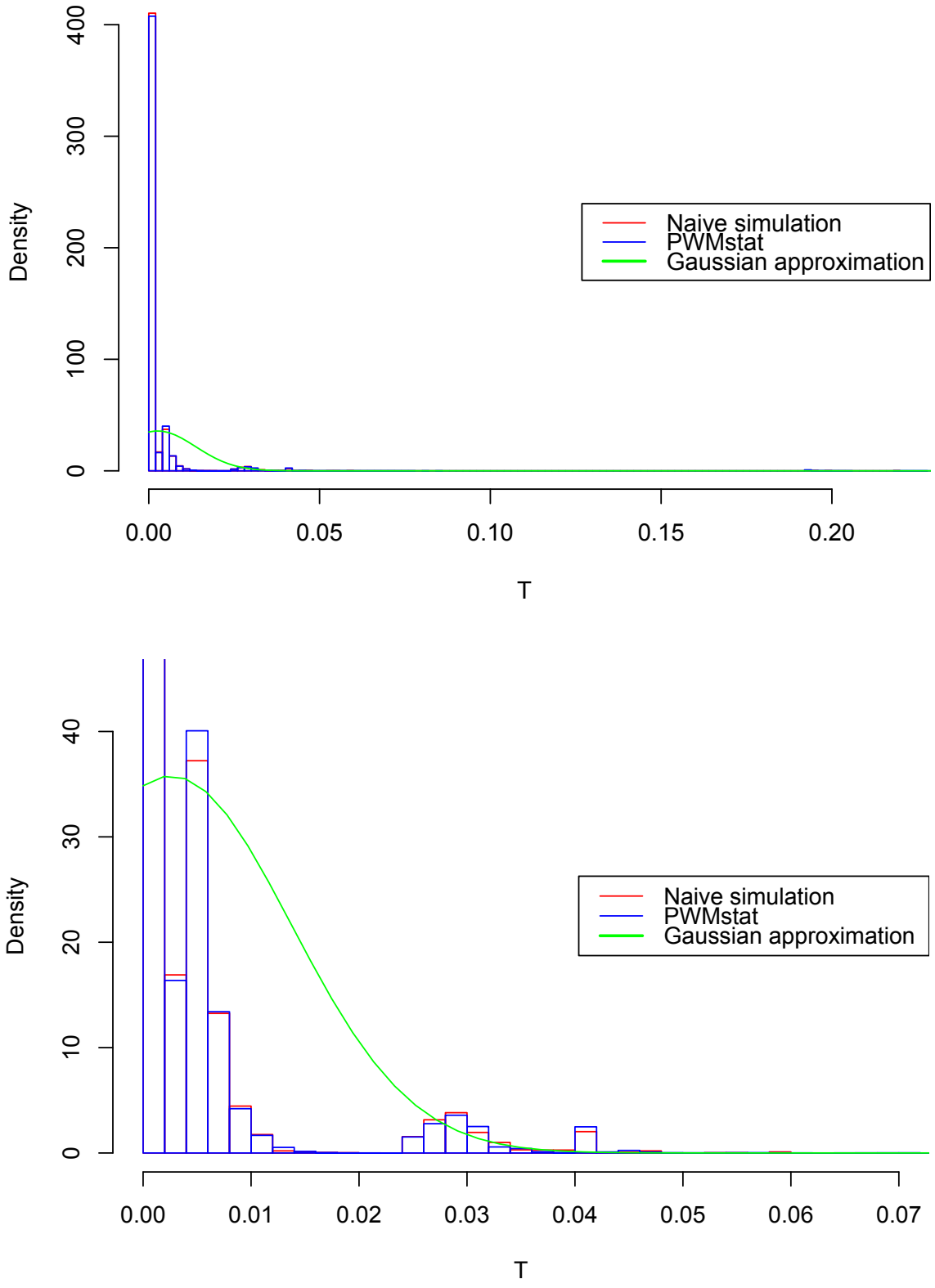



**Figure 3.** Distribution of the weighted count $T$ for the NBS motif of *B. subtilis* : empirical distribution (red), compound Poisson approximation (blue) and Gaussian distribution (green). The bottom figure is a zoom of the upper figure.

### 5.2 Comparison of PWMstat and Bioprospector Score

Bioprospector is a discovery motif tool which scans a DNA sequence and provides a list a PWMs ranked according to a specific score function; this score is supposed to reflect the enrichment significance of the PWM. We then wanted to compare how Bioprospector and PWMstat rank the PWMs. We did not use the *B. subtilis* sequence from the previous section because the NBS motif seems so frequent that Bioprospector does not find other PWM sufficiently different from NBS. We have then generated a Bernoulli sequence of one million letters and nucleotide frequency $p(a) = p(c) = 0.27$ and $p(g) = p(t) = 0.23$ (same composition than the sequence used in the previous section) and run Bioprospector to find enriched 12 bp-long motifs. Among the 20 highest scored motifs, only 12 motifs have been considered as different PWMs (PWMs with very close probabilities are considered identical), and we computed their $p$-value thanks to our approach (PWMstat with $10^7$ sampling). Fig. 4 compares the Bioprospector score of each motif ($x$-axis) with the probit transformation of the PWMstat $p$-value ($y$-axis); clearly, both scores are absolutely not correlated meaning that the Bioprospector score is not equivalent to the significance of the weighed count of the PWM as defined in (4). For example, the 11-th highest scored motif according to BioProspector is considered a thousand times more exceptionnal according to the PWMstat score than the top scored motif from BioProspector, therefore illustrating that BioProspector scores do not generally reflect well the enrichment significance of a PWM as defined by the weighted count $T$.

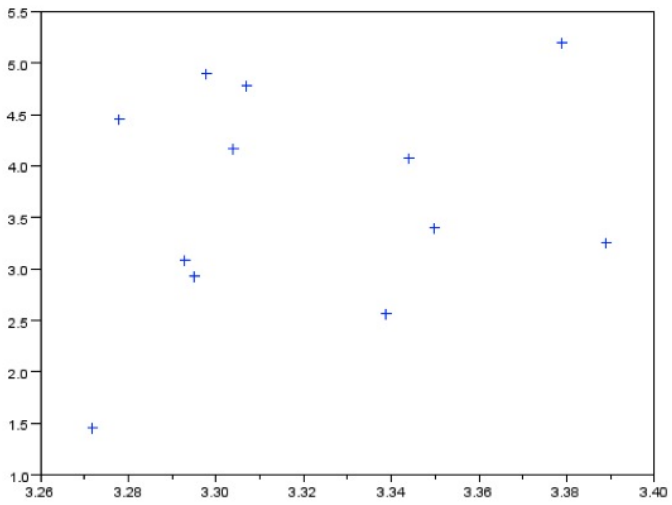



**Figure 4.** Enrichment scores calculated by Bioprospector ($x$-axis) and PWMstat ($y$-axis) for the 12 different PWMs with highest bioprospector scores. The score for PWMstat is simply the probit transformation of the $p$-value namely the quantile of the standard Gaussian distribution associated to a right tail equal to the $p$-value.

## 6 Conclusion

Computing $p$-values to evaluate if occurrences of PWMs in a DNA sequence are significant or simply due by chance is a problem faced by all motif discovery tools but also everytime one wants to scan a DNA sequence for known PWMs.

Results from word count statistics cannot be applied as soon as the number of compatible words is too large, which is often the case with binding motifs. Classicaly, existing methods study the occurrences of compatible words whose weight exceed a certain threshold, whatever the value of their weight. To take this information into account, we then proposed to measure the enrichment/frequency of a PWM by its weighted count, which is a natural generalization of a word count.

We show that the distribution of the weighted count can be well approximated by a compound Poisson distribution and we provide an algorithm to sample this distribution. The efficiency of the algorithm allows to get the empirical distribution of the weighted count with a great accuracy (one million sampling in 100 seconds for the 14bps-long NBS motif) leading to the $p$-value.

Our method could then be used by, or like a complement of, motif discovery tools which rank candidate PWMs according to specific scores. As shown in our study, such scores do not really reflect the significance of the PWM weighted count.

## Acknowledgements

We thank Valentin Belissen, Joao-Felipe Cabral-Moraes, Benjamin Dartevelle and Romain Faugeroux with whom this work has been initiated in the framework of their Projet Scientifique Collectif X2010.

## References

[1] Arratia, R., Goldstein, L. and Gordon, L. Poisson approximation and the Chen-Stein method. *Statistical Science*. 5: 403–434, 1990.

[2] Bailey, T.L. and Elkan, C. Fitting a misture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*. 2: 28–36, 1994.

[3] Hartmann, H., Guthöhrlein, E.W., Siebert, M., Luehr, S. and Söding, J. P-value-based regulatory motif discovery using positional weight matrices. *Genome Res.*. 23: 181–194, 2013.

[4] Liu, X., Brutlag, D.L. and Liu, J.S. Bioprospector: discovering conserved DNA motifs in upstream regulatory regions of cp-expressed genes. *Pac. Symp. Biocomput.*. 127–138, 2001.

[5] Manke, T., Roider, H. G. and Vingron, M. Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS computational biolog* 4(3) e1000039, 2008.

[6] Pape, U.J., Rahmann, S., Sun, F. and Vingron, M. Compound Poisson approximation of the number of occurrences of a position frequency matrix (PFM) on both strands. *J. Comput. Biol.*. 15: 547–564, 2008.

[7] Robin, S., Rodolphe, F. and Schbath, S. *DNA, Words and Models*, Cambridge University Press, 2005, English version of *ADN, mots et modèles*, BELIN 2003.

[8] Robin, S. and Schbath, S. Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.* 8: 349–359, 2001.

[9] Roider, H. G., Kanhere, A., Manke, T. and Vingron, M. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* 23(2): 134–141, 2007.

[10] Roquain, E. and Schbath, S. Improved compound Poisson approximation for the number of occurrences of multiple words in a stationary Markov chain, *Advances in Applied Probability*, 39: 128–140, 2007.

[11] Schbath, S. Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics*. 1: 1–16, 1995.

[12] Touzet, H. and Varré, J.-S. Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms for Molecular Biology*. 2:15, 2007.

[13] Wu, L.J. and Errington, J. Coordination of cell division and chromosome segregation by a nucleoid occlusion protein in Bacillus subtilis. *Cell*. 117: 915–925, 2004.

[14] Wu, L.J., Ishikawa, S., Yoshikazu Kawai, Y., Oshima, T., Ogasawara, N. and Errington, J. Noc protein binds to specific DNA sequences to coordinate cell division with chromosome segregation. *EMBO J.*. 28(13): 1940–1952, 2009.

# Graph analysis of chromatin conformation data in relation with the human replication program

Rasha E. BOULOS[1], Alain ARNEODO[1], Pablo JENSEN[1,2] and Benjamin AUDIT[1]

[1] Université de Lyon, F-69000 Lyon and Laboratoire de Physique, ENS de Lyon, CNRS, F-69007 Lyon, France.
[2] IXXI, Rhône Alpes Institute of Complex Systems, F-69007 Lyon, France.
prenom.nom@ens-lyon.fr

**Abstract** *We use graph theory to analyze chromatin interaction (Hi-C) data in the human genome. We show that "master" replication origins that border megabase-sized U-shaped replication timing domains are vertices of high degree-, betweenness-, and eigenvector- centralities in the chromatin interaction network. These early replication initiation zones, specified by a* $\sim 200$ *kb wide region of accessible open chromatin permissive to transcription, are found to form long-distance interconnected hubs of DNA interactions both within chromosomes and in between different chromosomes. The concomitant local enrichment in open chromatin marks and insulator elements suggests that these master replication origins are at the heart of a high-order epigenetically controlled 3D organization of the human genome into replication foci.*

**Keywords** Hi-C data, chromatin interaction graph, replication domains, nodes centralities.

It is increasingly recognized that the 3D dynamical architecture of eukaryotic genomes plays an important role in the regulation of nuclear functions [1,2,3,4,5,6,7,8]. The recent development of the chromosome conformation capture (3C) technology [9] and its high-throughput extensions [10,11] including Hi-C [12], has led to rapid advances in the study of the so-called tertiary chromatin structure in higher eukaryotes [10,11,12,13,14,15,16,17,18,19,20]. Hi-C techniques provide genome-wide quantitative measurement of the interaction frequency between pairs of loci. Because of cell-to-cell variation in chromatin folding, these chromatin interaction data have been considered as a *proxy* of some average 3D spatial proximity over a conformational ensemble. Pioneering works using 3C technologies [10,11,21] have shown that long-range looping interactions over tens or hundreds of kilobases are involved in gene regulation, for example between promoters, enhancers and insulators. Hi-C technology has revealed that drosophila [13,14] and mammalian [16,17] genomes are further organized in megabase-sized topological domains where long-range interactions can bring together highly transcribed genes offering a mechanism to explain the formation of transcription factories [1,2,3,4,5,6,7,8,22,23]. Hi-C data have also been compared to genome-wide mean replication timing (MRT) data of related cell lines [24,25] and consistent with the original Hi-C analysis of the human genome [12], some dichotomic picture has been proposed where early and late replication loci occur in seperated compartments of open and closed chromatin respectively. Interestingly, when using a wavelet-based multi-scale analysis of MRT data [26,27,28,29] for seven human cell types, about half of the genome was shown to be divided in megabase-sized U-shaped MRT domains that likely correspond to Hi-C topological domains of self-interacting chromatin [30]. Significant overlap is observed between the MRT U-domains of different cell types and also with germline replication domains previously defined as exhibiting a N-shaped nucleotide compositional profile [31,32,33,34,35,36]. From the demonstration that the average fork polarity is directly reflected by both the compositional skew and the derivative of the MRT profile [30,37,38], it has been argued that the fact that the MRT derivative displays a N-shape in U-domains sustains the existence of megabase-sized gradients of replication fork polarity in somatic and germline human cells [27,28,29,30,31]. When investigating the large scale organization of human genes inside these replication domains, a remarkable organization has been revealed [33,35]; in particular highly expressed genes in a given cell type are over-represented close to the corresponding U/N-domain borders [39]. When further mapping chromatin mark data inside U/N-domains, the "master" replication origins at U/N-domain borders were shown to be specified by a region ($\sim 200$ kb) of open and transcriptionally active chromatin [30,40] that is significantly enriched in insulator DNA-binding proteins (CTCF) [30]. A recent high-resolution 4C study dedicated to the analysis of the interaction of some selected

U/N-domain borders with the rest of the human genome [20] has confirmed that these early-initiation zones play a major role in the chromatin tertiary structure. Our aim here is to use graph theory [41] to objectively quantify the importance of the "master" replication origins at U/N-domains borders in the genome-wide intra-chromosomal and interchromosomal Hi-C chromatin network.

Graphs [41] have become extremely useful as a representation of a wide variety of complex systems in social sciences, biology, computer sciences and engineering [42,43,44,45]. An undirected graph $G = (V, E)$ consists of a finite set $V$ of *vertices* ($n = |V|$) and a finite set $E \subseteq V \times V$ of *edges* $e_{ij}$ ($m = |E|$) with associated weights $w_{ij} > 0$. Note that unweighted graphs correspond to $w_{ij} = 1$. It is often useful to consider a matrix representation of a graph. The *adjacency matrix* $\mathcal{A}$ is a $n \times n$ square matrix whose entry $a_{ij}(i, j = 1, .., n)$ is equal to the weight $w_{ij}$ and zero otherwise. Here we define the length of a path $(e_{i_1 i_2}, e_{i_2 i_3}, ..., e_{i_{k-1} i_k})$ between vertices $v_{i_1}$ and $v_{i_k}$ as $L = \sum_{l=1}^{k-1} \frac{1}{w_{i_l i_{l+1}}}$. To identify and quantify the vertices that occupy critical positions in a network, *centrality measures* have been proposed including the degree-, betweenness- and eigenvector centralities [42,43,44,45]. They can all be derived from the adjacency matrix. An intuitive ranking of the vertices of a graph is obtained by sorting them according to the total weight of their incident edges. The corresponding *degree centrality* $C_d$ is defined as [46]: $C_d(v_i) = \sum_{j=1}^{n} a_{ij}$. $C_d$ is a local centrality measure since it takes only into account the local structure (vertex neighborhood) of the graph. *Betweenness centrality* $C_b$ measures the extent to which a vertex lies between other vertices on their *geodesic* paths defined as the path of shortest lenght $L$ [46]: $C_b(v_i) = \sum_{\substack{j,k=1 \\ j \neq k}}^{n} \frac{\sigma_{jk}(v_i)}{\sigma_{jk}}$, where $\sigma_{jk}$ is the number of shortest paths connecting $v_j$ and $v_k$, while $\sigma_{jk}(v_i)$ is the number of shortest paths connecting $v_j$ and $v_k$ and passing through $v_i$. A vertex with high betweenness centrality can potentially influence the spread of information through the graph by facilitating or hindering the communication between others. To distinguish vertices that are linked to well-connected vertices and so may influence many others in the graph either directly or indirectly through their connections, Bonacich [47] has proposed the following definition of the *spectral centrality*: $C_\lambda(v_i) = \frac{1}{\lambda} \sum_{j=1}^{n} a_{ij} C_\lambda(v_j)$. This leads to the eigenvector computation $\lambda C_\lambda = \mathcal{A} C_\lambda$ and the eigenvector of the largest eigenvalue ($\lambda_{max}$) is the *eigenvector centrality* $C_{\lambda_{max}}$.

To illustrate the results of the centrality measure analysis, we use the open source Gephi software [48], a tool to visualize and manipulate large graphs. For the dynamical evolution of our graph, we choose the special force-directed algorithm called *Force Atlas 2*. This algorithm amounts to simulate a physical system of particles (vertices) distributed in a plane (2D): vertices repulse each other like magnets while edges attract the vertices they connect like springs. These forces create a dynamics that converges to a balanced final state that is expected to help the interpretation of the data. Note that graphs do not always converge to the same final configuration since many local stationary states may exist. In our numerical simulations, random 2D positioning of the weighted graph vertices is systematically used as initial condition.

We have performed our analysis of Hi-C experimental data [12] by mainly focusing on the intra- and inter-chromosomal contact maps obtained in the human erythroid cell line K562 (100 kb resolution maps from the GEO website: GSE18199_binned_heatmaps). These Hi-C contact maps are positively defined and symmetric and so can be represented and analyzed using graph theory [49]. We consider the Hi-C contact matrix as the adjacency matrix $\mathcal{A}$ of a weighted graph $G$, where the vertices $v_i$ are the 100 kb DNA loci and the edges $e_{ij}$ are weighted according to the number of Hi-C binary interactions. Because the number of intrachromosome interactions decreases very fast when increasing the separation $s$ between the loci ($\sim s^{-1}$) [12,49], the weighted network amounts to focus on interactions between loci separated by short genomic distances ($\lesssim 10$ Mb) over which contact probabilities are the highest. Alternatively, the non weighted version of the network takes equally into account short-range and long-range interactions within a chromosome. In this case, we optionally remove from the data all binary interactions that are present only once ($t = 1$) or twice ($t = 2$) as some of these may well be attributed to experimental noise (t=0 corresponds to no thresholding). In Fig. 1 is shown a Hi-C contact matrix (Fig. 1(b)) corresponding to intrachromosome interactions on a 12 Mb fragment of human chromosome 10 where 4 MRT U-domains were identified (Fig. 1(a)) in K562 [27,28,29,30]. As sketched by the dashed squares in Fig. 1(b), these 4 U-domains likely correspond to 4 matrix-square blocks of enriched interactions. This observation suggests that MRT U-domains correspond to some spatial compartmentalization into self-interacting structural chromatin units where the bordering early initiation zones prevent cross-talk between

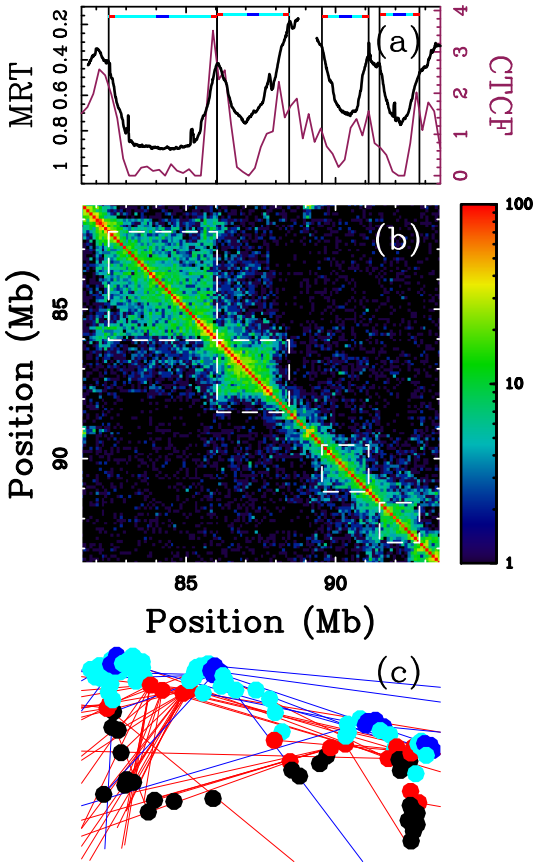



**Figure 1.** (a) MRT profile (black curve) [30,50] from early 0 to late 1 along a 12 Mb fragment of human chromosome 10 in K562. The horizontal colored bars correspond to the identified 4 MRT U-domains (red: 200 kb borders, dark blue: 400 kb center, light blue: interior). CTCF enrichment profile (purple curve) (ENCODE Release 3, Mar 2010) [51]. (b) Corresponding intrachromosome Hi-C contact matrix [12]. Each pixel represents the total number of interactions between pairs of 100 kb loci. The dashed squares delimit interactions within the 4 U-domains. (c) Stationary configuration obtained for this 12 Mb fragment when running the Gephi software to visualize the chromosome 10 interaction graph (Fig. 3(a)). Vertices are colored according to their position relative to U-domains: border (red), center (dark blue), interior (light blue), exterior (black). The represented edges correspond to connections between respectively U-domain borders (red) and centers (dark blue) with their neighbors distant from more than 4 Mb. The contact threshold t=2 (see text).

these domains [30]. To quantify the importance of these U-domains borders in the Hi-C contact interaction graph, we perform a statistical analysis over the 876 U-domains ($\leq$ 3 Mb) identified in the X and 22 human autosomes in K562 [30]. We also consider 140 additional "split U/N-domains" of size $\geq$ 3 Mb whose borders have similar gene organization and chromatin structure as U/N-domain borders [26].

In Fig. 2(a-c) are reported the results of the calculation of the 3 previously defined centralities as a function of the distance to the closest replication domain border. For both the unweighted and weighted graph descriptions, these borders are local maxima of the 3 centralities confirming that they correspond to critical vertices in the Hi-C interaction network. In Fig. 2(a), the degree centrality $C_d$ for the unweighted graph decreases almost linearly when moving inside the replication domains. This decrease of $C_d$ mainly results from the fact that borders have more long-range connections than centers (up to 2-fold for t=2) as quantified in (Fig. 2(d,e)). This is illustrated in Fig. 1(c) where we clearly see on a 12 Mb DNA fragment that the edges between vertices distant by more than 4 Mb and that are incident to MRT domain borders (red) are significantly more numerous than the ones that emanate from domain centers (blue). At smaller genomic distances ($\leq$ 2 Mb), the graph is almost fully connected (Fig. 2(d)) so that center loci are as much connected as borders. When considering the weighted graph description, the relative decrease of $C_d$ from border to center is much more pronounced than before (Fig. 2(a)) indicating that borders have a much higher intrachromosome contact frequency than centers and this regardless of the intragenomic distance ($>$ or $\leq$ 4 Mb). In Fig. 2(b), the betweenness centrality $C_b$

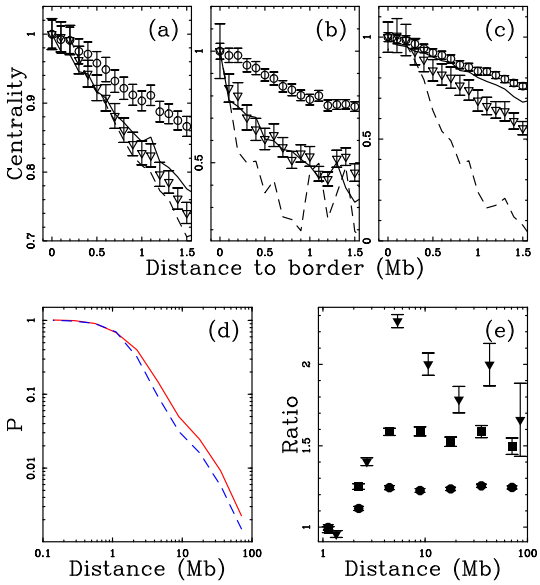



**Figure 2.** (a) Degree-, (b) betweenness-, (c) eigenvector intrachromosomic centralities *vs* the distance to the closest border of the 1016 K562 replication domains [30]: unweighted graph (◯), weighted graph for all edges (△), edges connecting loci distant by more than 4 Mb (—), or less than 4 Mb (- - -). $C_d$, $C_b$ and $C_\lambda$ are represented relative to the value found at MRT domain borders. (d) Probability distribution function of the genomic distances between MRT domain borders (red) and centers (blue) and their neighboring vertices for contact threshold $t = 1$. (e) Ratio of the numbers of incident edges to MRT domain borders and centers *vs* their intragenomic distances for different contact thresholds t=0 (●), 1(■) and 2 (▼). To better discriminate centers from borders, only the MRT domains longer than 2 Mb are taken into consideration in (d) and (e).

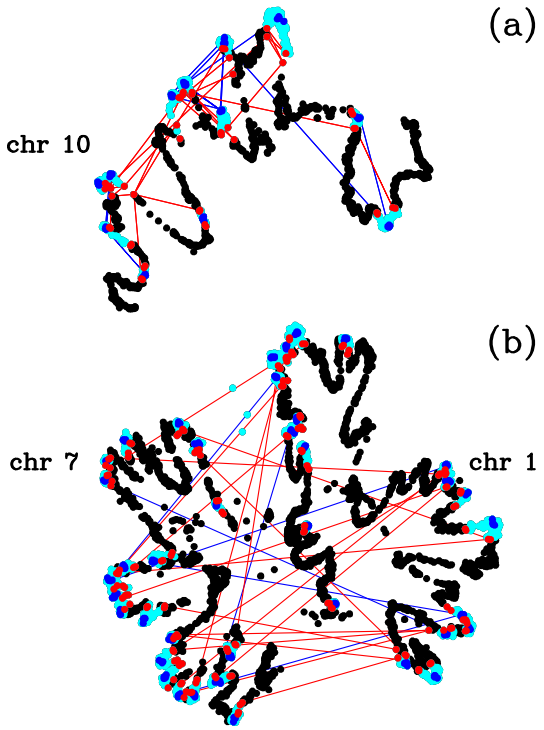



**Figure 3.** Stationary configuration obtained after running the Gephi software from the K562 HiC interaction matrices: (a) chromosome 10, (b) chromosomes 1 and 7. The color dots have the same meaning as in Fig. 1(c) for U-domains longer than 2 Mb. The color edges correspond to connections between two domain borders (red) or two centers (blue) for genomic distances larger than 4 Mb in (a) and for interchromosomal connection in (b). The contact threshold t=2. Similar illustrative pictures are obtained when considering other pairs of human chromosomes.

| t | border/border | center/center | Ratio |
|---|---|---|---|
| 0 | $3,44.10^{-2}$ | $2,57.10^{-2}$ | $1,34$ |
| 1 | $1,02.10^{-3}$ | $0,65.10^{-3}$ | $1,57$ |
| 2 | $4,20.10^{-5}$ | $1,23.10^{-5}$ | $3,41$ |

**Table 1.** Probability of interchromosome interaction between borders (resp. centers) of replication domains longer than 2 Mb for different contact thresholds t=0, 1 and 2. The last column represents the ratio between the border/border and center/center probabilities.

is also shown to decrease faster for the weighted graph. This quantifies the role of these replication domain borders as "hubs" in the chromatin interaction network. These hubs are not only important actors in mediating long-range interactions ($> 4$ Mb), they also play a fundamental role at short distance ($\leq 4$ Mb) where the very sharp decrease of $C_b$ observed in Fig. 2(b) enlightens the insulator properties of these domain borders that prevent cross-talk between neighboring domains likely establishing self-interacting independent expression domains (see the dashed squares in Fig. 1(b)) [30]. In Fig. 2(c), the eigenvector centrality $C_\lambda$ also decreases from MRT domain borders to center. The comparison of the relative decay obtained with the weighted and unweighted graphs further strengthens the fundamental organizing role of the MRT domain borders. They are hubs that predominantly interact with other hubs and especially at short distance ($\leq 4$ Mb) as a possible signature of 3D looping proximity [20,21]. The preferential long-range interactions between distal hubs is visualized in Fig. 3(a). Replication domain borders are found closer to each other in the stationary graph configuration, unlike centers that mainly interact at short distance and are found at the periphery of this stationary worm like final configuration (Fig. 1(c)).

We then extend our analysis to interchromosome interactions using the unweighted graph approach. In Table 1 are reported the probabilities of interchromosome interactions for MRT domain borders and centers deduced from the numbers of incident edges to borders and centers for different contact thresholds t=0, 1 and 2. Similarly to long-range intrachromosome connections, we find that borders have a higher probability of interchromosome interactions. Several sources of possible biases such as local GC content have been described to affect Hi-C experimental procedures [52,53]. Here, we confirmed that the results reported in Table 1 remain unchanged when applying the same analysis to the Hi-C data normalized using the procedure proposed in Ref. [53]. Hence, MRT domain borders are highly interconnected hubs, as illustrated in Fig. 3(b) for the graph of interactions between chromosomes 1 and 7. Clearly this stationary configuration sheds light on the importance of the red connections between MRT domain borders of these two chromosomes as compared to the few blue connections between domain centers. This explains that as regard to the peripheral disposition of the blue loci (centers) in both chromosomes, the red loci (borders) have a tendency to be closer together in each chromosome locally forming a horseshoe-like pattern, but also to be closer to their partners on the other chromosomes.

To summarize, we have used graph theory to quantify the predominant role played by the "early" replicating initiation zone that border replication U/N-domains in the human intrachromosome and interchromosome Hi-C chromatin network. Foregoing work showed that these replication MRT domain borders are zones of open and transcriptionally active chromatin, hypersensitive to DNase I, enriched in CpG islands and Pol II ChIP-Seq tags [30]. In that work, these regions also appeared to be significantly enriched in insulator binding protein CTCF as illustrated in Fig. 1(a); this observation is consistent with the conclusion of a recent study [49] concerning the potential role of CTCF in the formation of chromosomal hubs of interactions across chromosomes. Thereby, these so-called "master" replication initiation zones [40] are not only barrier elements that delimit self-interacting topological domains of independant expression and duplication [30], but they also mediate long-range interactions among distant DNA elements within chromosomes and in between chromosomes. It remains to be determined if this property is specific for those U/N-domain borders or if it is shared by other regions caracteristic of the large-scale organization of the human genome. Our results provide also some comprehensive understanding of the recent observation that long-range chromatin interaction loci are enriched in evolutionary breakpoints [54]. As reported in previous work [55], the replication U/N-domain borders are significantly enriched in mammalian evolutionary breakpoints suggesting that evolution typically shuffled these topological and functional domains rather than breaking them. These domains were actually shown to be conserved in mammals [32,33,34,35,36], confirming that the specific spatio-temporal replication program

underlying U/N-domains is likely to be a major determinant of genome 3D architecture and evolution.

This work was supported bu the ANR (REFOPOL, ANR 10 BLAN 1615).

## References

[1] P. R. Cook, The organization of replication and transcription, *Science*, 284(5421):1790–1795, 1999.

[2] T. Cremer and C. Cremer, Chromosome territories, nuclear architecture and gene regulation in mammalian cells, *Nat. Rev. Genet.*, 2(4):292–301, 2001.

[3] R. Berezney, Regulating the mammalian genome: the role of nuclear architecture., *Adv. Enzyme Regul.*, 42:39–52, 2002.

[4] N. Gilbert, S. Gilchrist and W. A. Bickmore, Chromatin organization in the mammalian nucleus, *Int. Rev. Cytol.*, 242:283–336, 2005.

[5] T. Misteli, Beyond the sequence: cellular organization of genome function., *Cell*, 128(4):787–800, 2007.

[6] T. Sexton, H. Schober, P. Fraser and S. M. Gasser, Gene regulation through nuclear organization, *Nat. Struct. Mol. Biol.*, 14(11):1049–1055, 2007.

[7] M. R. Branco and A. Pombo, Chromosome organization: new facts, new models., *Trends Cell. Biol.*, 17(3):127–134, 2007.

[8] P. Fraser and W. Bickmore, Nuclear organization of the genome and the potential for gene regulation., *Nature*, 447(7143):413–417, 2007.

[9] J. Dekker, K. Rippe, M. Dekker and N. Kleckner, Capturing chromosome conformation, *Science*, 295:1306–1311, 2002.

[10] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel and W. de Laat, Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C), *Nat. Genet.*, 38(11):1348–1354, 2006.

[11] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green and J. Dekker, Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements, *Genome Res.*, 16(10):1299–1309, 2006.

[12] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander and J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, *Science*, 326(5950):289–293, 2009.

[13] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay and G. Cavalli, Three-dimensional folding and functional organization principles of the drosophila genome., *Cell*, 148(3):458–472, 2012.

[14] C. Hou, L. Li, Z. S. Qin and V. G. Corces, Gene density, transcription, and insulators contribute to the partition of the drosophila genome into physical domains., *Mol. Cell*, 48(3):471–484, 2012.

[15] Y. Zhang, R. P. McCord, Y.-J. Ho, B. R. Lajoie, D. G. Hildebrand, A. C. Simon, M. S. Becker, F. W. Alt and J. Dekker, Spatial organization of the mouse genome and its role in recurrent chromosomal translocations., *Cell*, 148(5):908–921, 2012.

[16] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber and L. Chen, Genome architectures revealed by tethered chromosome conformation capture and population-based modeling., *Nat. Biotechnol.*, 30(1):90–98, 2012.

[17] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu and B. Ren, Topological domains in mammalian genomes identified by analysis of chromatin interactions., *Nature*, 485(7398):376–380, 2012.

[18] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Bl?thgen, J. Dekker and E. Heard, Spatial partitioning of the regulatory landscape of the X-inactivation centre., *Nature*, 485(7398):381–385, 2012.

[19] S. I. Takebayashi, V. Dileep, T. Ryba, J. H. Dennis and D. M. Gilbert, Chromatin-interaction compartment switch at developmentally regulated chromosomal domains reveals an unusual principle of chromatin folding., *Proc. Natl. Acad. Sci. USA*, 109(31):12574–12579, 2012.

[20] B. Moindrot, B. Audit, P. Klous, A. Baker, C. Thermes, W. de Laat, P. Bouvet, F. Mongelard and A. Arneodo, 3D chromatin conformation correlates with replication timing and is conserved in resting cells., *Nucleic Acids Res.*, 40(19):9470–9481, 2012.

[21] E. Splinter and W. de Laat, The complex transcription regulatory landscape of our genome: control in three dimensions., *EMBO J*, 30(21):4345–4355, 2011.

[22] D. R. F. Carter, C. Eskiw and P. R. Cook, Transcription factories, *Biochem. Soc. Trans.*, 36(Pt 4):585–589, 2008.

[23] S. Schoenfelder, T. Sexton, L. Chakalova, N. F. Cope, A. Horton, S. Andrews, S. Kurukuti, J. A. Mitchell, D. Umlauf, D. S. Dimitrova, C. H. Eskiw, Y. Luo, C.-L. Wei, Y. Ruan, J. J. Bieker and P. Fraser, Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells., *Nat. Genet.*, 42(1):53–61, 2010.

[24] E. Yaffe, S. Farkash-Amar, A. Polten, Z. Yakhini, A. Tanay and I. Simon, Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture, *PLoS Genet.*, 6:e1001011, 2010.

[25] T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton and D. M. Gilbert, Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types, *Genome Res.*, 20(6):761–770, 2010.

[26] A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. d'Aubenton-Carafa and C. Thermes, Multi-scale coding of genomic information: From DNA sequence to genome structure and function, *Phys. Rep.*, 498:45–188, 2011.

[27] B. Audit, A. Baker, C.-L. Chen, A. Rappailles, G. Guilbaud, H. Julienne, A. Goldar, Y. d'Aubenton Carafa, O. Hyrien, C. Thermes and A. Arneodo, Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm., *Nat. Protoc.*, 8(1):98–110, 2013.

[28] B. Audit, L. Zaghloul, A. Baker, A. Arneodo, C.-L. Chen, Y. d'Aubenton Carafa and C. Thermes, Megabase replication domains along the human genome: relation to chromatin structure and genome organisation., *Subcell. Biochem.*, 61:57–80, 2012.

[29] G. Guilbaud, A. Rappailles, A. Baker, C.-L. Chen, A. Arneodo, A. Goldar, Y. d'Aubenton-Carafa, C. Thermes, B. Audit and O. Hyrien, Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome, *PLoS Comput. Biol.*, 7(12):e1002322, 2011.

[30] A. Baker, B. Audit, C.-L. Chen, B. Moindrot, A. Leleu, G. Guilbaud, A. Rappailles, C. Vaillant, A. Goldar, F. Mongelard, Y. d'Aubenton Carafa, O. Hyrien, C. Thermes and A. Arneodo, Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines, *PLoS Comput. Biol.*, 8(4):e1002443, 2012.

[31] E.-B. Brodie of Brodie, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes and A. Arneodo, From DNA sequence analysis to modeling replication in the human genome, *Phys. Rev. Lett.*, 94(24):248103, 2005.

[32] M. Touchon, S. Nicolay, B. Audit, E.-B. Brodie of Brodie, Y. d'Aubenton-Carafa, A. Arneodo and C. Thermes, Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins, *Proc. Natl. Acad. Sci. USA*, 102(28):9836–9841, 2005.

[33] M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo and C. Thermes, Human gene organization driven by the coordination of replication and transcription, *Genome Res.*, 17(9):1278–1285, 2007.

[34] B. Audit, S. Nicolay, M. Huvet, M. Touchon, Y. d'Aubenton Carafa, C. Thermes and A. Arneodo, DNA replication timing data corroborate in silico human replication origin predictions, *Phys. Rev. Lett.*, 99(24):248102, 2007.

[35] A. Baker, S. Nicolay, L. Zaghloul, Y. d'Aubenton-Carafa, C. Thermes, B. Audit and A. Arneodo, Wavelet-based method to disentangle transcription- and replication-associated strand asymmetries in mammalian genomes, *Appl. Comput. Harmon. Anal.*, 28:150–170, 2010.

[36] C.-L. Chen, L. Duquenne, B. Audit, G. Guilbaud, A. Rappailles, A. Baker, M. Huvet, Y. d'Aubenton Carafa, O. Hyrien, A. Arneodo and C. Thermes, Replication-associated mutational asymmetry in the human genome, *Mol. Biol. Evol.*, 28(8):2327–2337, 2011.

[37] A. Baker, H. Julienne, C. L. Chen, B. Audit, Y. d'Aubenton Carafa, C. Thermes and A. Arneodo, Linking the DNA strand asymmetry to the spatio-temporal replication program. I. about the role of the replication fork polarity in genome evolution., *Eur. Phys. J. E*, 35(9):92, 2012.

[38] A. Baker, C. L. Chen, H. Julienne, B. Audit, Y. d'Aubenton Carafa, C. Thermes and A. Arneodo, Linking the DNA strand asymmetry to the spatio-temporal replication program: II. accounting for neighbor-dependent substitution rates., *Eur. Phys. J. E*, 35(11):123, 2012.

[39] L. Zaghloul, A. Baker, B. Audit and A. Arneodo, Gene organization inside replication domains in mammalian genomes, *C. R. Mécanique*, 340:745–757, 2012.

[40] B. Audit, L. Zaghloul, C. Vaillant, G. Chevereau, Y. d'Aubenton-Carafa, C. Thermes and A. Arneodo, Open chromatin encoded in DNA sequence is the signature of "master" replication origins in human cells, *Nucleic Acids Res.*, 37(18):6064–6075, 2009.

[41] B. Bollobas, *Modern Graph Theory*, Springer, New York, USA, 1998.

[42] S. Wasserman and K. Faust, *Social Network Analysis*, Cambridge University Press, Cambridge, 1994.

[43] M. Newman, The structure and function of complex networks, *SIAM Rev.*, 45(2):167–256, 2003.

[44] A.-L. Barabási and Z. N. Oltvai, Network biology: understanding the cell's functional organization., *Nat. Rev. Genet.*, 5(2):101–113, 2004.

[45] S. Boccaletti, V. Latora, Y. Moreno, M. Charez and D. Hwang, Complex networks: Structure and dynamics, *Phys. Rep.*, 424:175, 2006.

[46] L. Freeman, Centrality in social networks conceptual clarification, *Social Networks*, 1:215, 1979.

[47] P. Bonacich, Factoring and weighting approaches to status scores and clique identification, *Journal of Mathematical Sociology*, 2:113, 1972.

[48] M. Bastian, S. Heymann and M. Jacomy, Gephi: An open source software for exploring and manipulating networks, Third International AAAI Conference on Weblogs and Social Media. Available at: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/ view/154, 2009.

[49] M. Botta, S. Haider, I. X. Y. Leung, P. Lio and J. Mozziconacci, Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide., *Mol. Syst. Biol.*, 6:426, 2010.

[50] R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler and J. A. Stamatoyannopoulos, Sequencing newly replicated DNA reveals widespread plasticity in human replication timing, *Proc. Natl. Acad. Sci. USA*, 107(1):139–144, 2010.

[51] The ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, 447(7146):799–816, 2007.

[52] E. Yaffe and A. Tanay, Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture, *Nat. Genet.*, 43(11):1059–1065, 2011.

[53] A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul and J. Mozziconacci, Normalization of a chromosomal contact map, *BMC Genomics*, 13:436, 2012.

[54] A. S. Véron, C. Lemaitre, C. Gautier, V. Lacroix and M.-F. Sagot, Close 3d proximity of evolutionary breakpoints argues for the notion of spatial synteny., *BMC Genomics*, 12:303, 2011.

[55] C. Lemaitre, L. Zaghloul, M.-F. Sagot, C. Gautier, A. Arneodo, E. Tannier and B. Audit, Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation, *BMC Genomics*, 10:335, 2009.

# Session 2 : Analyse de séquences

## Conférence invitée

RODERIC GUIGO

Center for Genomic Regulation
Universitat Pompeu Fabra
BARCELONA

# The landscape of transcription in human cells

The unfolding of the instructions encoded in the genome is triggered by the transcription of DNA into RNA, and the subsequent processing of the resulting primary RNA transcripts into functional mature RNAs. RNA is thus the first phenotype of the genome, mediating all other phenotypic changes at the organism level caused by changes in the DNA sequence. While current technology is too primitive to provide accurate measurements of the RNA content of the cell, the recent development of Massively Parallel Sequencing Instruments has dramatically increased the resolution with which we can monitor cellular RNA. Using these instruments, the ENCODE project has surveyed the RNA content of multiple cell lines and subcellular compartments. The results of these surveys underscore pervasive transription, as well as great RNA heterogeneity between and within cells. Comparison of RNA surveys with other genome wide epigenetic surveys — such as those of binding sites for Transcription Factors, or of Histone modifications — reveals a very tightly coupling between the different pathways involved in RNA processing, transcription and splicing in particular. Overall, the recent large scale transcriptome and epigenome surveys reveal that large portions of the genome exhibits some sort of biochemical activity. What fraction of this activity can be associated to biological function remains an open question

# The genome of the medieval Black Death agent (extended abstract)

Ashok Rajaraman[1,2], Eric Tannier[3,4], Cedric Chauve[1,5]

[1] Department of Mathematics, Simon Fraser University, V5A 1S6 Burnaby (BC), Canada
{arajaram,cedric.chauve}@sfu.ca
[2] International Graduate Training Center in Mathematical Biology, Pacific Institute for Mathematical Sciences, Vancouver (BC), Canada
[3] INRIA Grenoble Rhône-Alpes, F-38334 Montbonnot, France
eric.tannier@inria.fr
[4] Université de Lyon 1, Laboratoire de Biométrie et Biologie Évolutive, CNRS UMR5558 F-69622 Villeurbanne, France
[5] LaBRI, Université Bordeaux I, 33405 Talence, France

**Abstract** *The genome of a 650 year old* Yersinia pestis *bacteria, responsible for the medieval Black Death, was recently sequenced and assembled into 2,105 contigs from the main chromosome. According to the point mutation record, the medieval bacteria could be an ancestor of most* Yersinia pestis *extant species, which opens the way to reconstructing the organization of these contigs using a comparative approach. We show that recent computational paleogenomics methods, aiming at reconstructing the organization of ancestral genomes from the comparison of extant genomes, can be used to correct, order and complete the contig set of the Black Death agent genome. The obtained sequence suggests that a burst of mobile elements insertions predated the Black Death, leading to an exceptional genome plasticity and increase in rearrangement rate.*

**Keywords** Paleogenomics, computational biology, genome assembly, pathogens

## Le génome de la bactérie responsable de la Peste Noire

**Résumé** *Récemment, le génome d'une souche de la bactérie* Yersinia pestis *vieille de 650 ans a té séquencée et assemblée en 2,105 contigs issus de son chromosome. Cette bactérie médiévale semble être l'ancêtre de la plupart des souches actuelles de* Yersinia pestis*, ce qui permet d'appliquer une approche comparative pour assembler ces contigs en scaffolds. En utilisant des méthodes et principes récemment développés pour la reconstruction de l'organisation de génomes anciens à partir de la comparaison de génomes existants, nous corrigeons, organisons et complétons les contigs de l'agent de la Peste Noire. L'analyse de la séquence ainsi obtenuesuggère que de nombreuses insertions d'éléments mobiles ont participé à l'émergence d'un génome execeptionellement dynamique et à une augmentation du taux de réarrangements.*

**Mots-clés** Paléogénomique, bioinformatique, assemblage de génomes, pathogènes.

## 1 Introduction

The plague has long been among the most feared human diseases [10], due to dramatic pandemics such as the *Black Death* which ravaged Europe in the late middle-ages. Recently Bos *et al.* [8] were able to sequence the whole genome of the Black Death agent, and concluded that it was an ancestor of most extant strains of the human pathogen *Yersinia pestis* (see also [44]). The sequence extracted from the oral metagenome of one individual was assembled using Velvet [48], into approximately 130,000 contigs, including 2,105 contigs of length $\geq$ 500bp from the main chromosome, with similarities with some *Yersinia* extant genomes. This first sequencing of the chromosome of an extinct prokaryote helped to understand the causes of the Black Death pandemic [8, 37, 47]. However, the assembled 2,105 contigs cover only 85% of the expected length of the ancestral chromosome, and their organization along this ancestral chromosome is unknown, keeping out of reach a detailed genome-scale study of the evolution of the structural organization of *Yersinia* genomes, whose impact on pathogenicity is still an important open question [11].

Current assembly methodologies can hardly be applied to fully assemble and finish an ancient genome, aside of short molecules such as plasmids [44] and organelle genomes [36]. Indeed, existing scaffolding methods, aimed at ordering and orienting the contigs, and estimating the lengths of inter-contig gaps, rely on additional data such as mate-pair reads with mixed insert sizes [2, 22, 40, 41, 49], optical or physical maps [27] or comparison with one or several closely related genomes [24, 42]. However, due to the decay and fragmentation of ancient DNA, reads from ancient genomes are in general short, and optical maps or mate-pair libraries with long inserts can not be obtained [1]. This leaves the comparative approach as the only possibility to handle sequence data such as the one of the Black Death agent. The usual setting of the comparative approach involves the comparison of the contigs with one, or a few, closely related genomes, either genome sequence or maps [5, 7, 24, 34] or protein sequences [42]. However, to the best of our knowledge, none of these methods is intended to be applied on the genome of an internal node of a given phylogeny.

We describe a comparative approach to scaffold an ancient genome, and apply it to the medieval plague agent. The ancestral Black Death agent is indeed related to a dozen of descendants (from the *Yersinia pestis* clade) and close outgroups (from the *Yersinia pestis* and *Yersinia pseudotuberculosis* clades), whose phylogeny, taken from Bos *et al.* [8], is shown on Fig. 1.

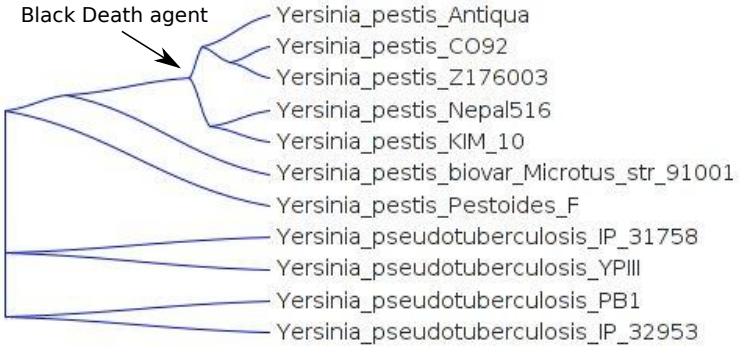



**Figure 1.** Phylogeny of the used extant genomes and position of the reconstructed one.

There has been a recent flurry of ancestral genome organization reconstruction methods, complementing classical methods for reconstructing ancestral genome sequence [6, 19, 26] and gene content [15, 16, 45]. They have been used for reconstructing ancestral genomes of bacterias [21,46], animals [1,9,14,29,31,33,35,38,39], plants [32, 43], yeasts [4, 13, 23] or protists [28]. Recent developments provide exact and fast algorithms that handle sequence duplications, repeats, diverse types of genome rearrangements and chromosome structures [3, 25, 30].

We show here that this corpus of methods is efficient and versatile enough to be integrated into a comparative scaffolding framework for ancient bacterial genomes, and we illustrate this claim with a complete assembly of the medieval Black Death agent chromosome. Starting from the alignment of the contigs assembled by Bos *et al.* [8] with the sequences of extant *Yersinia* genomes, we compute a single circular scaffold containing the ordered and oriented sequences from the whole set of contigs, completed by estimations of the sequences located between consecutive contigs (gaps). Additionally, we correct some contigs initially assembled by Bos *et al.* by identifying probable *chimeric*, *redundant* or *duplicated* contigs. The chromosome structure we observe is distant from every extant genome, explaining the difficulty of the assembly process with a single reference genome. We annotate and analyze the ancestral chromosome, pointing at a probable replication origin, predicting the positions of insertion sequences (IS) and detecting the numerous inversions that separate it from the considered extant genomes. We provide evidence that the speciation between the *Yersinia pestis* and *Yersinia pseudotuberculosis* clades was characterized by a burst of insertion of IS elements in the *Yersinia pestis* genomes, concomitant with an increase rate of genome rearrangements, which breakpoints positions are also correlated with IS.

1. The reads of the current data set are in fact single reads.

## 2 Results

The main result of our work is a completely assembled chromosome sequence of the Black Death agent genome. To obtain it, we followed a generic procedure for reconstructing an ancestral genome organization [4, 14, 25, 29, 31], which comprises four phases: (1) extracting homologous families of ancestral and extant genome markers, (2) computing putative linkage between ancestral markers, (3) combining the set of ancestral linkages into a circular sequence of ancestral markers, (4) inferring inter-marker gap sequences. We provide only a sketch of the implementation in this extended abstract; full details will appear elsewhere.

*Families of homologous segments.* We aligned the ancestral contigs against 11 fully assembled genomes of *Yersinia* strains. Several contigs were not aligned over their full length on every genome because of rearrangements. So we did cut such contigs into segments, such that every segment is aligned over its full length and no pair of genomic segment defined by two different alignments overlap (they are either disjoint or confounded). This clusters ancestral and extant genome segments into 2,619 homologous families. Each family contains one or several ancestral contig segments (called *ancestral markers* from now), and zero, one or several genome segments (called *extant markers* from now)from each extant species.

All sequences from a single family are assumed to be homologous, that is, to share a common ancestor and to have evolved from this ancestor through speciations, duplications, losses or transfers. We do not have phylogenetic trees for the families that would allow us to detect those events and derive a marker content (i.e. copy number) [45]. Yet some ancestral markers correspond to repeated sequences that were present at several loci of the ancestral genome, while some of them contain ancestral segments from several different contigs. We used phyletic profiles [15, 16] to determine the number of occurrences of every ancestral marker, namely the ancestral marker content of this ancestral genome. We computed this ancestral content for each family by using a parsimony approach that minimizes the number of gains and losses of markers along the species tree for each family. This allows to associate to each family a *multiplicity*, *i.e.* its expected number of occurrences in the ancestral chromosome; 20 families out of 2,619 have a multiplicity greater than 1.

The amount of DNA encoded by the markers, when multiplicity is accounted for, is 3,846,866bp of ancestral DNA, while the initial contigs encode 4,013,159bp. This initial loss of sequenced ancient DNA will be compensated by filling the gaps between the different pieces of the segmented contigs.

*Computing putative linkages between ancestral markers.* We computed sets of ancestral markers that are believed to be consecutive in the ancestral chromosome. We call them *intervals* of ancestral markers, if they contain more than two markers and *adjacencies* if they concern only two markers. We followed a Dollo parsimony principle [14] to infer putative ancestral linkages: a group of ancestral markers is deemed to be contiguous in the ancestral genome if markers from the same families are contiguous in at least two extant genomes whose evolutionary path on the species phylogeny contains the ancestor of interest (here the Black Death agent). All 2,637 putative adjacencies obtained in this way are then weighted according to their phylogenetic conservation, using a recursive formula inspired from the Fitch-Hartigan principle [4, 14, 29].

*Combining the set of ancestral linkages into a circular sequence of ancestral markers.* The set of putative ancestral adjacencies is not compatible with a circular chromosomal structure, due to possible converging genome rearrangements, for example. Indeed some markers may be involved in too many adjacencies. However, discarding 6 adjacencies out of the 2,637 putative ancestral adjacencies was enough to obtain a set of maximal cumulative weight that can be ordered circularly. They were found implementing a fast and exact "circularization" method based on matching techniques in graphs [30], that aims at findiigna maximal subset of adjacencies that can be arranged into a set of linear and circular segments while satisfying the constraints given by their multiplicities.

Adjacencies alone are compatible with many circular orders due to repeated ancestral markers forming tangles in the adjacency graph [2, 24]. To address this issue, intervals of size greater than two spanning groups of markers with multiplicity greater than 1 were used, as illustrated in Fig. 2, resulting in an ordering of the markers into three large scaffolds.

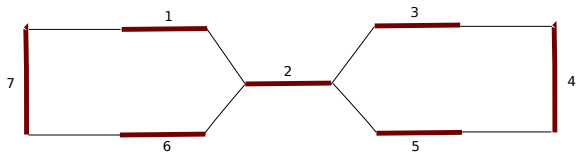



**Figure 2.** Illustration of the ambiguity in ordering ancestral markers with multiplicities greater than 1 and of the use of intervals to address it. Here is a toy example where we have markers $1, \ldots, 7$, drawn with bold red segments, and adjacencies between their extremities, drawn with thin black lines. Assume every marker has multiplicity 1 except marker 2, which has multiplicity 2. Then every marker extremity has as many adjacencies as its multiplicity predicts. But there are several possible circular orderings or these markers according to these adjacencies: 1,2,3,4,5,2,6,7, or 1,2,5,4,3,2,6,7. Suppose we have in addition size three intervals, and among them we find $\{1, 2, 3\}$ or $\{2, 5, 6\}$. Then only the first ordering is compatible. In our data set, intervals up to size 6 were sufficient to completely clarify the adjacency signal.

We then joined the extremities of these three scaffolds to form a circular chromosome by choosing, among the six possible configurations, the only one supported by some extant genomes. This resulted into a complete circular ordering of ancestral markers, where each ancestral marker appears exactly as many times as it is expected from its multiplicity.

*Correcting the initial contigs.* In the resulting ordering, each occurrence of an ancestral marker corresponds to one or several segments of the initial contigs. The ordering of these segments is mostly compatible with the initial contigs. We found only one *chimeric* contig (see Fig. 3), split into two non-adjacent markers in the ancestral genome organization. None of the extant occurrences from the two families are adjacent in extant genomes, pointing to either an assembly error during the initial contig construction, or a derived rearrangement in the ancient genome, which would be interesting since Bos *et al* [8] did not find such a mutation looking at nucleotide substitutions. Note that the length filtering applied onto families after the contig segmentation step can lead to an underestimation of the number of chimeric contigs: if part of a contig has length less than the threshold, it is discarded and the contig is not detected as chimeric. Also four contigs segments were found to be *duplicated*: a large part ($> 500$bp) of each is probably present in more than one occurrence in the ancestral genome, while the initial assembly predicted only one occurrence. Finally, 63 contigs have a sequence which is found, up to very small variations, inside another contig while their number of extant occurrences suggest they have multiplicity one, so we believe they are *redundant*.

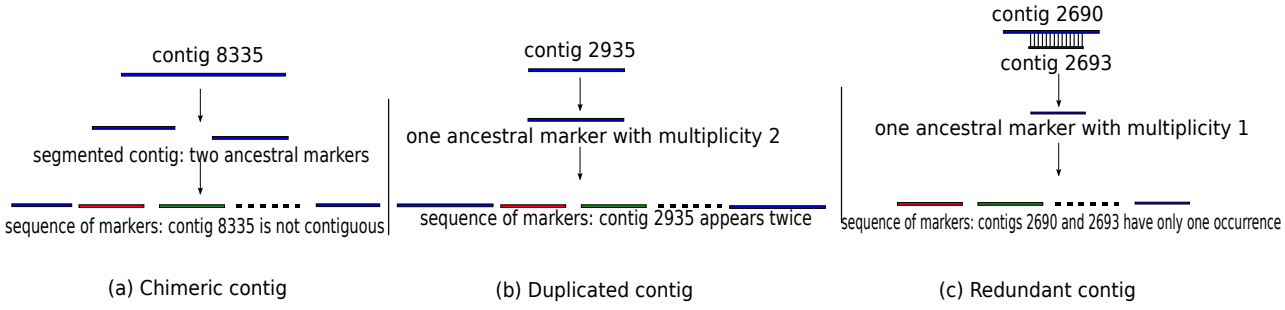



**Figure 3.** Contig correction: (a) the contig is cut during the segmentation procedure, but not joined during the marker ordering; (b) the contig is found to have two occurrences in the marker ordering; (c) two contigs contain the same DNA sequence and this sequence is predicted to have only one occurrence in the marker ordering.

*Estimating ancestral gaps sequences.* We completed this assembly by estimating the sequences located in *ancestral gaps*, *i.e.* between pairs of ancestral markers consecutive in the circular ordering. For this we first

estimated a length interval for each ancestral gap: a length is said to be *supported* for an ancestral gap if there are two gaps in extant genomes, in two species whose evolutionary path contains the ancestor of interest, with such a length. The length interval of a gap is defined by the minimum and maximum supported length for this gap. For 24 gaps we found no supported length, so we took the minimum and maximum gap length of extant sequences in the species where the markers are consecutive. Then for each ancestral gap, we aligned all extant gaps which lengths fall in the ancestral gap length interval. We then constructed an ancestral sequence from each alignment by a reconstruction method for ancestral discrete character s implementing the Fitch algorithm [20].

This resulted into an ancestral genome sequence of length 4,586,856 showing that 739,990bp were added to the ancestral markers sequences by this finishing step. Only 1 gap was not assigned a sequence by this method.

*Analysis of the reconstructed ancestor.* We took advantage of reconstructing the full chromosome of the Black Death agent to analyze its structure and evolution at the whole-genome scale.

We traced the GC-skew with SeqinR [12] from a CDS annotation by Glimmer (Fig. 4(b)) to predict the position of the replication origin. We slipped the reconstructed ancient genome sequence such that the putative replication origin (the maximum value in the cumulative GC-skew plot) has position 0 and we aligned the ancient chromosome with the chromosome of the CO92 strain. We obtained the dotplot represented in Fig. 4(a) that shows the highly repeated nature of both genomes, and the rearrangements that have happened along the lineage from the ancestor to the CO92 strain.

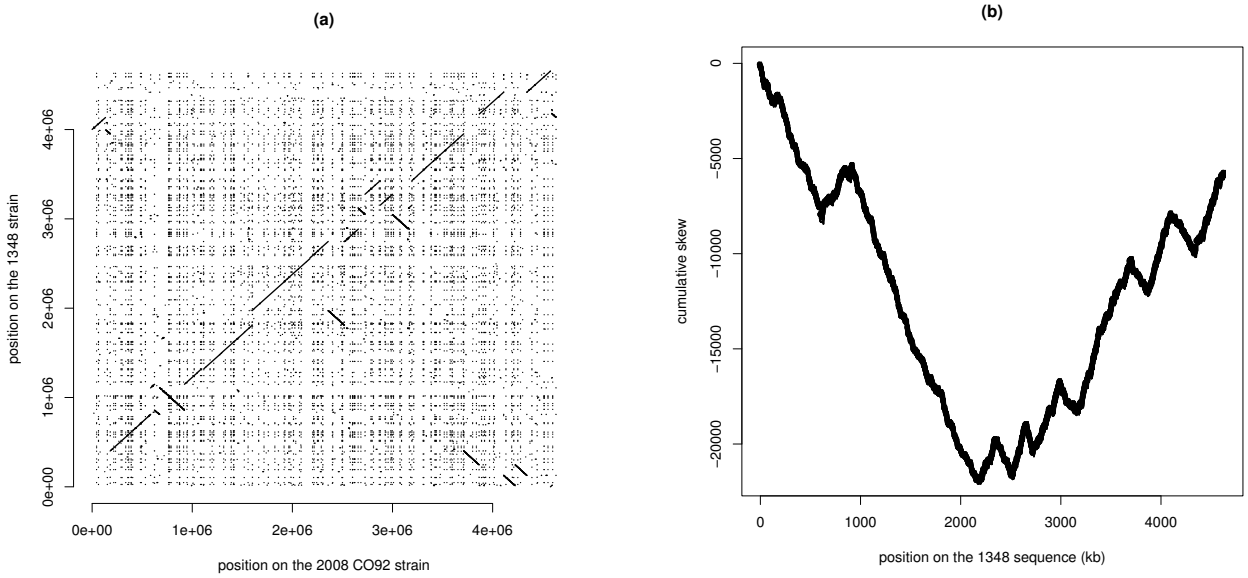



**Figure 4.** (a) Dotplot of all Megablast alignments of the medieval sequence against the CO92 extant strain. The highly repetitive nature of both genomes appear, as well as the inversions that happened in the CO92 history, several of them being symmetric around the origin of replication. (b) Cumulative skew shows a probable position for the replication origin (for which we chose position 0), as well as the rearrangements which tend to blur the skew signal.

We mapped IS elements onto the reconstructed ancestral chromosome, based on a conservative analysis of their patterns of presence in extant markers and gaps: an ancestral gap is assigned an IS if one of its occurrences in the descendants genomes is of length exactly the minimum length of the ancestral gap and contains an annotated IS; we focused on gaps as no extant marker does contain an annotated IS. This resulted in 94 ancestral gaps containing IS. We confirmed this comparative annotation with an automatic annotation. Our analysis also shows that a large part of these IS (at least 57) were already present in the last common ancestor of all *Yersinia pestis* strains, while they are almost completely absent from the genomes of the considered outgroups from the *Yersinia pseudotuberculosis* clade.

We also analyzed the genome rearrangements between the ancestral sequence and extant genomes by sampling inversion scenarios between the ancestral genome and the extant genomes (see Fig. 5). There are 8-9 inversions between the *Yersinia pseudotuberculosis* strains and the medieval genome, and 9-22 inversions when compared to (thought evolutionarily closer) *Yersinia pestis* strains. As noticed by Darling et al [17], we can also observe that inversion breakpoints are not randomly distributed and used: highly used ones are concentrated in one third of the chromosome, around its replication origin. Most inversions are symmetrical around the origin. The positions of the inversion breakpoints are also highly correlated with IS, as remarked earlier [18]: 76 out of the 118 mapped breakpoints are close ($< 1000$bp distant) to some predicted IS, while this number drops to 39 for uniformly sampled random coordinates (p-value $< 10^{-3}$). Rearrangements are very numerous in all *pestis* branches, strongly suggesting that they could be driven by the IS.

| | |
|---|---|
| Yersinia pestis biovar Microtus str 91001 | 22 |
| Yersinia pestis Pestoides F | 13 |
| Yersinia pseudotuberculosis IP 31758 | 9 |
| Yersinia pseudotuberculosis YPIII | 8 |
| Yersinia pseudotuberculosis PB1 | 9 |
| Yersinia pseudotuberculosis IP 32953 | 8 |
| Yersinia pestis Antiqua | 21-22 |
| Yersinia pestis CO92 | 12 |
| Yersinia pestis Z176003 | 13 |
| Yersinia pestis Nepal516 | 9 |
| Yersinia pestis KIM 10 | 9 |

**Figure 5.** Rearrangement distances between the extinct genome and the extant genomes. Two numbers mean that sampled scenarios have different length as we sample scenarios following a Bayesian posterior distribution of all scenarios, and not only the most parsimonious ones.

## 3 Discussion/Conclusion

The present work illustrates the potential of phylogenetic/comparative assembly methods to address the specific issues of ancient DNA assembly (single reads, fragmentation, ...). Our main result is a complete assembly of the chromosopme of a 650 years old bacteria, that opens the way to whole genome analysis of rearrangements and insertion dynamics among others.

The method we developed for this assembly relies on recent advances, both methodological and algorithmic, in reconstructing the organization of ancient genomes from the comparison of related extant genomes. We show here that such methods are generic enough to be also used with data acquired by sequencing of ancient DNA.

A crucial issue of such a method is its validation. In this extended abstract we do not develop this point but we are currently extensively testing our method on simulated data generated from *Yersinia* genomes.

We believe the methodological advances we present in this work complement the recent breakthrough in ancient DNA sequencing, at least for bacterial genomes, and suggest that integrating ancient genomes into comparative genomics is an ambitious but realistic goal for the next few years.

## Acknowledgments

This work was supported by NSERC Discovery Grant to C.C., a PIMS IGTC Fellowship to A.R. and ANR-10-BINF-01-01 Ancestrome to E.T. We are thankful to Laurent Duret, Vincent Daubin, Annie Chateau, Eric Rivals, Hendrik Poinar for useful discussions.

## References

[1] M. A. Alekseyev, P. A. Pevzner. Breakpoint graphs and ancestral genome reconstructions. *Genome Res*, 19(5):943–957, 2009.

[2] A. Bashir, A. Klammer, W. P. Robins, et al. A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotech*, 30(7):701–707, 2012.

[3] S. Bérard, C. Gallien, B. Boussau, G. J. Szollosi, Vincent Daubin, Eric Tannier. Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*, 28:i382–i388, 2012.

[4] D. Bertrand, Y. Gagnon, M. Blanchette, N. El-Mabrouk. Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss. In Mona Singh and Vincent Moulton, editors, *Algorithms in Bioinformatics, 10th International Workshop, WABI 2010, Liverpool, UK, September 6-8, 2010. Proceedings*, volume 6293 of *Lecture Notes in Bioinformatics*, pages 78– 89. Springer Verlag, 2010.

[5] D. Bertrand, M. Blanchette, N. El-Mabrouk. Genetic map refinement using a comparative genomic approach. *J Comput Biol*, 16(10):1475–1486, 2009.

[6] M. Blanchette, E. D. Green, W. Miller, D. Haussler. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*, 14(12):2412–2423, 2004.

[7] G. Blin, E. Blais, D. Hermelin, P. Guillon, M. Blanchette, N. El-Mabrouk. Gene maps linearization using genomic rearrangement distances. *J Comput Biol*, 14(4):394–407, 2007.

[8] K. I. Bos, V. J. Schuenemann, G. B. Golding, et al. A draft genome of yersinia pestis from victims of the black death. *Nature*, 478(7370):506–510, 2011.

[9] G. Bourque, P. A. Pevzner, G. Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res*, 14(4):507–516, 2004.

[10] A. Camus. *La peste*. Gallimard, 1947.

[11] P.S. Chain, E. Carniel, F.W. Larimer, et al. Insights into the evolution of yersinia pestis through whole-genome comparison with yersinia pseudotuberculosis. *Proc Natl Acad Sci U S A*, 101(38):13826–13831, 2004.

[12] D. Charif, J.R. Lobry. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York, 2007. ISBN : 978-3-540-35305-8.

[13] C. Chauve, H. Gavranovic, A. Ouangraoua, E. Tannier. Yeast ancestral genome reconstructions: the possibilities of computational methods II. *J Comput Biol*, 17(9):1097–1112, 2010.

[14] C. Chauve, E. Tannier. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput Biol*, 4(11):e1000234, 2008.

[15] O. Cohen, H. Ashkenazy, F. Belinky, D. Huchon, T. Pupko. Gloome: gain loss mapping engine. *Bioinformatics*, 26(22):2914–2915, 2010.

[16] M. Csurös. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912, 2010.

[17] A. E. Darling, I. Miklós, M. A. Ragan. Dynamics of genome rearrangement in bacterial populations. *PLoS Genet*, 4(7):e1000128, 2008.

[18] W. Deng, V. Burland, G. Plunkett, et al. Genome sequence of yersinia pestis kim. *J Bacteriol*, 184(16):4601–4611, 2002.

[19] A. B. Diallo, V. Makarenkov, M. Blanchette. Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*, 26(1):130–131, 2010.

[20] W. M. Fitch. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst Zool*, 20 (4):406–416, 1971.

[21] R. Fremez, T. Faraut, G. Fichant, J. Gouzy, Y. Quentin. Phylogenetic exploration of bacterial genomic rearrangements. *Bioinformatics*, 23(9):1172–1174, 2007.

[22] S. Gao, W.-K. Sung, N. Nagarajan. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *J Comput Biol*, 18:1681–1691, 2011.

[23] J. L. Gordon, K. P. Byrne, K. H. Wolfe. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome. *PLoS Genet*, 5(5):e1000485, 2009.

[24] P. Husemann, J. Stoye. Phylogenetic comparative assembly. *Algorithms Mol Biol*, 5:3, 2010.

[25] B. R. Jones, A. Rajaraman, E. Tannier, C. Chauve. ANGES: Reconstructing ancestral genomes maps. *Bioinformatics*, 28:2388–2390, 2012.

[26] D. A. Liberles, editor. *Ancestral Sequence Reconstruction*. Oxford University Press, 2007.

[27] H. C. Lin, S. Goldstein, L. Mendelowitz, S. Zhou, J. Wetzel, D. C. Schwartz, M. Pop. Agora: Assembly guided by optical restriction alignment. *BMC bioinformatics*, 13:189, 2012.

[28] J. Ma, A. Ratan, B. J. Shuh, L. Zhang, W. Miller, D. Haussler. Dupcar: reconstructing contiguous ancestral regions with duplications. *J Comput Biol*, 15:1007–1027, 2008.

[29] J. Ma, L. Zhang, B. B. Suh, B. J. Raney, R. C. Burhans, W. J. Kent, M. Blanchette, D. Haussler, W. Miller. Reconstructing contiguous regions of an ancestral genome. *Genome Res*, 16(12):1557–1565, 2006.

[30] J. Maňuch, M. Patterson, R. Wittler, C. Chauve, E. Tannier. Linearization of ancestral multichromosomal genomes. *BMC Bioinformatics*, 13(Suppl 19):S11, 2012.

[31] M. Muffato, A. Louis, C.-E. Poisnel, H. Roest Crollius. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, 26(8):1119–1121, 2010.

[32] F. Murat, J.-H. Xu, E. Tannier, M. Abrouk, N. Guilhot, C. Pont, J. Messing, J. Salse. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res*, 20(11):1545–1557, 2010.

[33] Y. Nakatani, H. Takeda, Y. Kohara, S. Morishita. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res*, 17(9):1254–1265, 2007.

[34] A. Muñoz, C.Zheng, Q. Zhu, V. A. Albert, S. Rounsley, D. Sankoff. Scaffold filling, contig fusion and comparative gene order inference. *BMC Bioinformatics*, 11:304, 2010.

[35] A. Ouangraoua, E. Tannier, C. Chauve. Reconstructing the architecture of the ancestral amniote genome. *Bioinformatics*, 27(19):2664–2671, 2011.

[36] J. L.Paijmans, M. T. Gilbert, M. Hofreiter. Mitogenomic analyses from ancient dna. *Mol Phylogenet Evol*, 1012. Epub ahead of print (Jun 15, 2012).

[37] J. Parkhill, B. W. Wren. Bacterial epidemiology and biology - lessons from genome sequencing. *Genome Biol*, 12:230, 2011.

[38] N. H. Putnam, T. Butts, D. E. K. Ferrier, et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198):1064–1071, 2008.

[39] N. H. Putnam, M. Srivastava, U. Hellsten, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, 317(5834):86–94, 2007.

[40] F. J. Ribeiro, D. Przybylski, S. Yin, et al. Finished bacterial genomes from shotgun sequence data. *Genome Res*, 22:2270–2277, 2012.

[41] L. Salmela, V. Mäkinen, N. Välimäki, J. Ylinen, E. Ukkonen. Fast scaffolding with small independent mixed integer programs. *Bioinformatics*, 27:3259–3265, 2011.

[42] S. L. Salzberg, D. D. Sommer, D. Puiu, V. T. lee. Gene-boosted assembly of a novel bacterial genome from very short reads. *PLoS Comput Biol*, 4:e1000186, 2008.

[43] D. Sankoff, C. Zheng, P. K. Wall, C. dePamphilis, J. Leebens-Mack, V. A. Albert. Towards improved reconstruction of ancestral gene order in angiosperm phylogeny. *J Comput Biol*, 16(10):1353–1367, 2009.

[44] V. J. Schuenemann, K. I. Bos, S. DeWitte, et al. Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of yersinia pestis from victims of the black death. *Proc Natl Acad Sci U S A*, 108:E746–E752, 2011.

[45] G. J. Szöllősi, B. Boussau, S. S. Abby, E. Tannier, V. Daubin. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci U S A*, 109:17513–17518, 2012.

[46] Y. Wang, W. Li, T. Zhang, C. Ding, Z. Lu, N. Long, J. P. Rose, B-C Wang, D. Lin. Reconstruction of ancient genome and gene order from complete microbial genome sequences. *J Theoret Biol*, 239:494–498, 2006.

[47] D. J. Wilson. Insights from genomics into bacterial pathogen populations. *PLoS Pathog*, 8:e1002874, 2012.

[48] D. R. Zerbino, E. Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*, 18(5):821–829, 2008.

[49] D.R. Zerbino, G. K. McEwen, E. H. Margulies, E. Birney. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One*, 4:e8407, 2009.

# Session 3A : Structure des protéines

# Visualize and study proteins via domain arrangement with DoMosaics

Andrew D. MOORE[1], Andreas HELD[1], Nicolas TERRAPON[1], January WEINER 3rd[2] and Erich BORNBERG-BAUER[1]

[1] Institute for Evolution and Biodiversity, Hüfferstrasse 1, 48147 Münster, Germany
{radmoore, n.terrapon, ebb}@uni-muenster.de
[2] Max Planck Institute for Infection Biology Chariteplatz 1, 10117 Berlin, Germany
january@mpiib-berlin.mpg.de

**Abstract** *DoMosaics is a tool designed for protein domain annotation and domain arrangement visualization and analysis. It simplifies the analysis of protein families by unifying disjunct procedures based on often inconvenient command-line based applications and complex analysis methodologies. First, DoMosaics provides easy, user interface based access to domain annotation services such as distant InterProScan queries or local HMMER installations. Second, it can be used to construct domain dotplots and adjacency-graphs, perform context-dependent identification of divergent domains, search for similar domain arrangements in a public protein database, and analyze the evolutionary history of protein families that is computing most parsimonious modular rearrangements along a phylogenetic tree. Furthermore, DoMosaics can create high quality, publication-ready images. DoMosaics is a Java application licensed under* Apache License 2.0. *Binaries and source code can be obtained from* www.domosaics.uni-muenster.de.

**Keywords** Java application, protein domains, modular evolution.

## Visualisation et étude des protéines par leur arrangement en domaines avec DoMosaics

**Résumé** *DoMosaics est un programme conçu pour l'annotation des domaines protéiques ainsi que la visualisation et l'analyse des arrangements en domaines. Il simplifie l'analyse des familles protéiques en unifiant des procédures indépendantes fréquemment basées sur des applications en ligne de commandes ou des méthodologies analytiques complexes. DoMosaics offre tout d'abord une interface utilisateur simplifiée pour accéder aux services d'annotation des domaines protéiques telles que des requêtes distantes au serveur InterproScan ou l'utilisation d'une installation locale du logiciel HMMER. Il inclut ensuite la reconstruction de dotplots et de graphes d'adjacence des domaines, l'identification des domaines divergents grâce au contexte en domaines, la recherche d'arrangements en domaines similaires dans une base publique de protéines et l'analyse de l'histoire évolutive d'une famille protéique grâce au calcul des réarrangements en domaines les plus parcimonieux pour un arbre phylogénétique donné. De plus, DoMosaics permet de créer des images de haute qualité, prêtes pour la publication. DoMosaics est une application Java sous licence* Apache License 2.0. *Les binaires et le code source sont disponibles sur* www.domosaics.uni-muenster.de.

**Mots-clés** Application Java, domaines protéiques, évolution modulaire.

## 1 Introduction

Les protéines n'évoluent pas seulement par des substitutions ponctuelles d'acides aminés mais également de façon modulaire par des réarrangements de plus longs segments. Ces segments coïncident fréquemment avec les domaines protéiques, les unités structurelles, fonctionnelles et évolutives des protéines. L'émergence de nouvelles protéines/fonctions est due en grande partie à l'apparition de nouveaux *arrangements en domaines* (*i.e.* séquence de domaines constitutifs de la protéine) [1]. La plupart des nouveaux arrangements en domaines qui apparaissent au cours de l'évolution peuvent être expliqués par des réarrangements simples [2]. Les mécanismes pressentis pour être les forces majeures en jeu sont les fusions/f ssions de protéines et les délétions terminales

de domaines [3]. Un certain nombre de bases de données dédiées aux domaines, telles que Pfam [4] ou la métabase Interpro [5], permettent de déterminer l'arrangement en domaines des protéines grâce à l'utilisation de librairies de modèles. Les outils existants qui exploitent l'arrangement en domaines des protéines se contentent généralement d'identifier un ensemble de protéines homologues (arrangements similaires) et sont souvent limités à une unique base de données de domaines (*e.g.* CDART [6], PfamAlyzer [7], WDAC [8], ArchSchema [9]). Seul le récent d-Omix [10] offre des outils simples pour accomplir une première étape de l'analyse de l'évolution modulaire des protéines. Il permet notamment d'utiliser plusieurs bases de données de domaines, de calculer et comparer des arbres basés sur les séquences ou les arrangements, et de visualiser sous forme de graphes des informations telles que la versatilité, l'abondance et les co-occurrences des domaines. La plupart de ces méthodes se présentent sous forme de serveurs web tels (CDART, WDAC, ArcShema) ou offrent des versions *stand-alone* qui nécessitent toutefois une connexion web (PfamAlyzer, d-Omix).

Nous décrivons ici un nouvel outil, DoMosaics, qui combine les différentes étapes pour l'analyse des arrangements en domaines en une ressource unique et pouvant s'affranchir du web. Outre l'annotation des domaines et la recherche d'arrangements similaires, DoMosaics propose une large gamme d'outils complémentaires pour faciliter l'étude de l'histoire évolutive des familles protéiques au niveau modulaire, dans la lignée de d-Omix, et l'addition de méthodes plus avancées telles que le calcul des réarrangements en domaines les plus parcimonieux en accord avec une phylogénie. De plus, DoMosaics offre un environnement graphique inédit permettant à l'utilisateur de manipuler le rendu des arrangements et de générer des images de haute qualité, prêtes pour la publication.

## 2 Application

Nous ne détaillerons pas ici l'ensemble des fonctionnalités de DoMosaics afin de nous concentrer sur les tâches les plus fréquentes de l'analyse évolutive des arrangements en domaines. Ce genre d'analyse démarre avec un sous-ensemble de séquences d'une famille protéique pour lesquelles on procède à l'annotation des domaines (section 2.1), à la recherche éventuelle de protéines ayant un arrangement similaire (section 2.2), à l'association des arrangements avec un arbre phylogénétique (section 2.3), et enfin à l'analyse de l'évolution modulaire de la famille (section 2.4).

### 2.1 Annotation des domaines

DoMosaics permet d'accéder à deux services classiques d'annotation en domaines : des requêtes distantes au serveur InterproScan [11] et l'utilisation d'une installation locale du logiciel HMMER3 [12]. Dans la première option, le programme InterproScan reproduit les méthodes/paramètres spécifiques à chacune des dix bases de données de domaines participant au consortium Interpro pour déterminer l'arrangement en domaines des protéines. Cette solution a l'avantage d'offrir une extension de perspectives par rapport à l'utilisation habituelle d'une unique ressource de domaines (généralement Pfam). De plus, InterproScan renvoie également les annotations fonctionnelles de la Gene Ontology [13] induites par la présence des domaines. La seconde option offre à l'utilisateur une interface à une version locale du programme HMMER3. Il suffit d'indiquer à DoMosaics l'emplacement des binaires `hmmpress` (compression de la librairie similaire au formatdb de blast) et `hmmscan` (pour la recherche de domaines) du package HMMER3, aucune *installation* physique d'HMMER n'est requise (*i.e.* le téléchargement des binaires suffit). Cette solution a l'avantage de permettre l'utilisation de n'importe quelle librairie de modèles au format HMMER3 (le plus répandu), soit téléchargée depuis une base de données de domaines publique telle que Pfam ou Superfamily [14], soit produite spécifiquement par l'utilisateur pour répondre à sa problématique. L'utilisateur est alors affranchi d'une quelconque présence ou limitation à une connexion au web.

Au-delà des services classiques d'annotation, DoMosaics implémente la détection de domaines divergents grâce au contexte en domaines, telle que décrite par la méthode CODD (*Co-Occurent Domain Detection*) [15]. CODD exploite une liste de paires de domaines précalculées (exhibant une co-occurrence significative) pour certifier les domaines Pfam qui ne satisfont pas les seuils de détection requis mais sont supportés par la présence d'autres domaines de l'arrangement.

Enf n, DoMosaics peut charger un f chier d'annotation de domaines. Le format requis s'inspire du format .xdom présenté pour la première fois dans le cadre de la base de données de domaines ProDom [16]. Ce format est proche du format fasta : une première ligne - avec un chevron ">" suivi du nom de la protéine - est suivie d'une ligne pour chaque domaine - où le domaine est décrit par ses positions de début, de f n et son nom séparés pas des tabulations-. Cela permet à l'utilisateur soit d'utiliser sa propre librairie de domaines (*e.g.* dans le cas de librairies privées comme chez CAZy [17]) ou de s'affranchir des méthodes d'annotation classiques pour générer sa propre composition en domaines de ses protéines.

Après l'annotation, DoMosaics permet une visualisation des arrangements unique et hautement personnalisable (*e.g.* formes, couleurs des domaines). L'utilisateur peut alors modif er cette vue de différentes façons, en jouant par exemple avec les seuils de détection, la visibilité des domaines recouvrants (en cas de hits multiples à une même position) ou l'alignement multiple des domaines répétés (automatiquement ou guidé manuellement par des mesures de similarités entre occurrences).

## 2.2 Recherche d'arrangements similaires

Une étape importante pour étudier l'évolution d'une famille de protéines consiste à rassembler l'ensemble de séquences homologues permettant de répondre à la problématique, en couvrant par exemple l'intégralité des paralogues d'une espèce ou un plus large échantillon taxonomique. Si ce travail n'a pas été effectué au préalable par l'utilisateur, DoMosaics propose une interface des méthodes RADS/RAMPAGE [18][19] pour retrouver et classif er les séquences homologues. Si de nombreuses méthodes, évoquées en introduction, s'attachent à la recherche d'homologues basés sur la composition en domaines, RADS est le premier algorithme calculant l'alignement entre deux arrangements en domaines. L'algorithme RAMPAGE s'appuie lui sur le résultat de RADS pour générer un alignement de séquences d'acides aminés : en exploitant l'alignement des domaines, il suit une procédure qui *divise pour mieux régner* et offre une solution sous-optimale de haute qualité avec un temps de calcul largement réduit. DoMosaics accède au serveur web de RADS/RAMPAGE pour récupèrer les arrangements similaires à un *arrangement-requête* parmi la base de données de séquences Uniprot [20].

## 2.3 Ajout d'une couche phylogénétique

Af n d'analyser l'histoire évolutive d'une famille protéique, la visualisation des arrangements peut être complétée par une représentation phylogénétique. L'arbre phylogénétique décrit les relations entre les arrangements en domaines (protéines) associés aux feuilles. L'utilisateur a la possibilité :

- soit d'utiliser DoMosaics pour créer un arbre grâce à l'une des deux approches proposées pour la reconstruction phylogénétique. La première, basée sur les séquences d'acides aminés, produit une requête au serveur web de l'EBI pour effectuer un alignement CLUSTALW [23] avant d'utiliser les fonctions de la librairie PAL (*Phylogenetic Analysis Library*) [24] pour calculer des arbres par UPGMA [25] ou Neighbor-Joining [26]. La seconde possibilité est de construire localement une topologie basée sur les arrangements en domaines en utilisant soit une distance d'édition soit une distance de Jaccard [27].
- soit de charger dans DoMosaics un arbre qu'il aura calculé par ses propres moyens. Cette solution est particulièrement adaptée pour un utilisateur souhaitant une topologie plus robuste à l'aide de programmes de phylogénie plus puissants mais qui nécessitent un plus grand temps de calcul ou de plus importantes ressources matérielles (*e.g.* avec morePhyML [21] ou Phylobayes [22]).

La visualisation des arbres phylogénétiques dans DoMosaics reproduit les fonctions décrites dans EPoS [28] pour la labélisation et la coloration des noeuds et des branches par exemple. Nous avons également intégré des méthodes spécif ques à l'analyse modulaire des protéines telle que l'agrégation taxonomique en fonction de l'identité ou du degré de similarité entre les arrangements.

## 2.4 Analyse des réarrangements

DoMosaics est également le premier logiciel à intégrer des fonctions simples aidant à élucider l'histoire évolutive de la famille protéique. Nous avons implémenté la reconstruction des arrangements ancestraux ainsi que le calcul des événements d'insertion et de délétion de domaines en accord avec un arbre phylogénétique

(*cf.* section 2.3). Pour cela, DoMosaics offre à l'utilisateur le choix entre deux principes de parcimonie suivant les algorithmes de Dollo [29] et Sankoff [30]. De plus, les coûts des événements d'insertion et de délétion sont paramétrables pour permettre à l'utilisateur de mieux évaluer la trajectoire évolutive des arrangements par l'introduction de connaissances *a priori* ou en testant la robustesse des résultats.

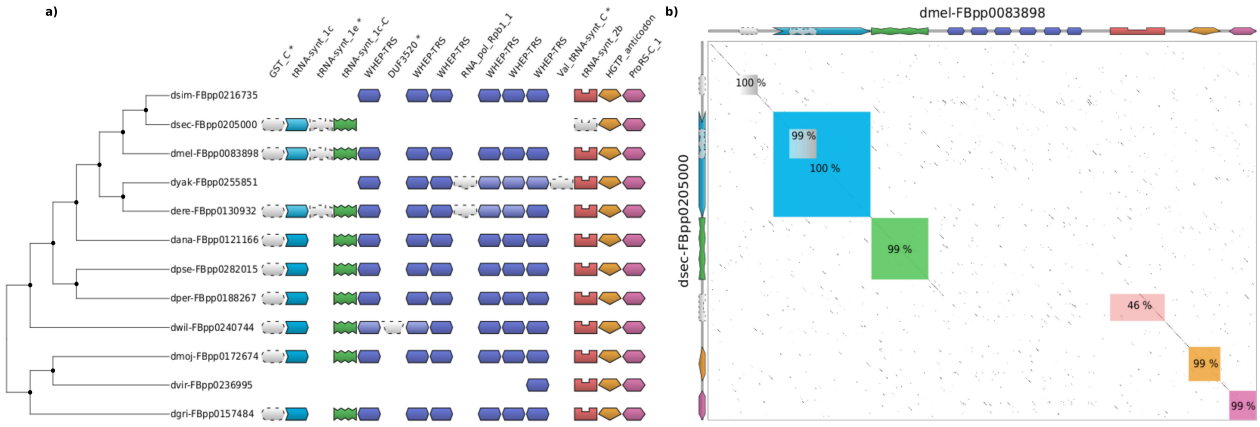



**Figure 1.** Analyse des aminoacyl-tRNA synthétases chez douze drosophiles grâce à DoMosaics. a) Arrangements en domaines associés à l'arbre des espèces. Les arrangements de domaines Pfam ont été determinés via HMMER avec une E-valeur de 2 pour appliquer la méthode CODD. Les domaines sont représentés par des boîtes non-proportionnelles, en couleur avec une bordure pleine pour les domaines qui satisfont les seuils Pfam, en blanc avec une bordure pointillée pour les domaines divergents récupérés par CODD. L'alignement multiple des domaines a été conduit manuellement en utilisant les pourcentages d'identité entre domaines. b) Dotplot des aminoacyl-tRNA synthétases de *D. melanogaster* et *D. sechellia* (indiqués en rouge en 1a). Les arrangements, représentés de manière proportionnelle sur les axes, sont utilisés pour ajouter au dotplot usuel une visualisation des paires de domaines de même type (rectangles de couleur) et une mesure de similarité pour chaque paire (en pourcentage).

## 3 Illustration

Pour illustrer l'interface de DoMosaics, nous présentons un bref exemple appliqué aux aminoacyl-tRNA synthétases chez douze drosophiles (Figure 1). Pour cela, nous avons extrait de Flybase [31] les orthologues de l'aminoacyl-tRNA synthétase de *D. melanogaster* (FBpp0083898) et chargé les séquences correspondantes dans un nouveau projet DoMosaics. L'annotation en domaines a été conduite avec une installation locale d'HMMER et la libraire de modèles Pfam. Le seuil de détection de domaines a été f xé à une E-valeur de 2, relâchant les seuils requis par Pfam (*gathering thresholds*) pour appliquer le f ltre CODD. CODD adjoint aux domaines satisfaisant les seuils Pfam (représentés en couleur) uniquement les domaines potentiels/divergents les plus vraisemblables (représentés en blanc) grâce au contexte en domaines. Nous avons ensuite chargé dans DoMosaics l'arbre phylogénétique communément admis des douze drosophiles séquencées.

Dans la Figure 1a, on observe des arrangements en domaines (non-proportionnels à la longueur des protéines) bien conservés pour huit des douze espèces. On trouve notamment un domaine N-terminal GST_C divergent (non-détecté par Pfam classique mais retrouvé chez Superfamily par exemple) présent dans ces huit protéines ainsi que chez *D. sechellia*. À l'opposé, on remarque une délétion N-terminale de plusieurs domaines, similaire chez *D. simulans* et *D. yakuba* et plus étendue chez *D. virilis*. Cependant, la non-monophylie de ces espèces suggère une convergence évolutive plutôt qu'un événement ayant eu lieu chez l'ancêtre commun : même le taxon rassemblant *D. simulans* et *D. yakuba* (sous-groupe *melanogaster*) inclut trois autres espèces dont aucune ne montre de perte de domaine N-terminale. Finalement, on note l'absence chez *D. sechellia* des répétitions de domaines WHEP-TRS et la divergence du domaine tRNAsynt_2b au centre tandis que les domaines terminaux sont bien conservés. Nous avons alors utilisé DoMosaics pour créer un dotplot des arrangements de *D. melanogaster* et *D. sechellia* (Figure 1b). Le dotplot révèle que la divergence tRNAsynt_2b est liée à l'absence de sa portion N-terminale (ligne interrompue sur la largeur du rectangle informatif) et est vraisemblablement liée à la

délétion de toute cette portion centrale (d'après la vue proportionnelle des arrangements). Si DoMosaics ne permet pas de résoudre l'origine de ces différences, entre réarrangements évolutifs réels ou artefacts d'annotations (provenant de domaines non-détectés ou de modèles de gènes erronés), il a permis une identif cation rapide, grâce à une visualisation ergonomique et des outils simples, des cas nécessitant une analyse plus détaillée.

## Remerciements

Les auteurs souhaitent remercier Stefan Rensing, Adam Sardar et Andreas Schüler pour avoir participer intensivement à tester les fonctionalités de DoMosaics et apporter des critiques constructives. Ce travail a été f nancé par le DFG (*Deutsche Forschungs Gemeinschaft*) bourse BO 2455/4-1 attribuée à EBB.

## Références

[1] E. Bornberg-Bauer, A.K. Huylmans, T. Sikosek, How do new proteins arise ? *Curr. Opin. Struct. Biol.*, 20(3) :390–396, 2010.

[2] A.K. Björklund, D. Ekman, S. Light, J. Frey-Skött and A. Elofsson, Domain rearrangements in protein evolution. *J Mol Biol.*, 353 :911-923, 2005.

[3] A.D. Moore, A.K. Björklund, D. Ekman, E. Bornberg-Bauer and A. Elofsson, Arrangements in the modular evolution of proteins. *Trends Biochem Sci.*, 33(9) :444-451, 2008.

[4] M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry and *et al.*, The Pfam protein families database. *Nucleic Acids Res.*, 40(Database issue) :D290-301, 2012.

[5] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T.K. Attwood and *et al.*, InterPro in 2011 : new developments in the family and domain prediction database. *Nucleic Acids Res.*, 40(Database issue) :D306-312 2012.

[6] L.J. Geer, M. Domrachev, D.J. Lipman and S.H. Bryant, CDART : protein homology by domain architecture. *Genome Res.*, 12 :1619-1623, 2002.

[7] V. Hollich and E.L. Sonnhammer, PfamAlyzer : domain-centric homology search. *Bioinformatics*, 23(24) :3382-3383, 2007.

[8] B. Lee and D. Lee, Protein comparison at the domain architecture level. *BMC Bioinf.*, 10 :S5, 2009.

[9] A.U. Tamuri and R.A. Laskowski, ArchSchema : a tool for interactive graphing of related Pfam domain architectures. *Bioinformatics*, 26(9) :1260-1261, 2010.

[10] D. Wichadakul, S. Numnark, S. Ingsriswang, d-Omix : a mixer of generic protein domain analysis tools. *Nucleic Acids Res.* 37(Web Server issue) :W417-421, 2009.

[11] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler and R. Lopez, InterProScan : protein domains identif er. *Nucleic Acids Res.*, 33(Web Server issue) :W116-120, 2005.

[12] S.R. Eddy, Accelerated prof le HMM searches. *PLoS Comp. Biol.*, 7 :e1002195, 2011.

[13] The Gene Ontology Consortium, Gene ontology : tool for the unif cation of biology. *Nat. Genet.*, 25(1) :25-29, 2000.

[14] D.A. de Lima Morais, H. Fang, O.J. Rackham, D. Wilson, R. Pethica, C. Chothia and J. Gough, SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.* 39(Database issue) :D427-434, 2011.

[15] N. Terrapon, O. Gascuel, É. Maréchal, and L. Bréhélin, Detection of new protein domains using co-occurrence : application to *Plasmodium falciparum*. *Bioinformatics*, 25(23) :3077-3083, 2009.

[16] J. Gouzy, P. Eugene, E.A. Greene, D. Kahn, and F. Corpet, XDOM, a graphical tool to analyse domain arrangements in any set of protein sequences. *Bioinformatics*, 13(6) :601-608, 2009.

[17] B.L. Cantarel, P.M. Coutinho, C. Rancurel, T. Bernard, V. Lombard and B. Henrissat, The Carbohydrate-Active enZymes database (CAZy) : an expert resource for glycogenomics. *Nucleic Acids Res.*, 37 :D233-D238, 2009.

[18] N. Terrapon, S. Grath, J. Weiner III, A.D. Moore and E. Bornberg-Bauer, Fast homology search using domain-architecture alignment. *JOBIM 2012*.

[19] N. Terrapon, J. Weiner III, S. Grath, A.D. Moore and E. Bornberg-Bauer, Rapid similarity search of proteins using alignments of domain arrangements. *Bioinformatics*, IN REVIEW, 2013.

[20] The UniProt Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 40 :D71-D75, 2012.

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

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

[23] J.D. Thompson, D.G. Higgins and T.J. Gibson, CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specif c gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22 :4673-4680, 1994.

[24] A. Drummond and K. Strimmer, PAL : an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics*, 17(7) :662–663, 2001.

[25] R.R. Sokal and C.D. Michener, A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* 38 :1409–1438, 1958.

[26] N. Saitou and M. Nei, The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4) :406-425, 1987.

[27] A. D. Moore and E. Bornberg-Bauer, *Protein Domains as Evolutionary Units*, John Wiley & Sons Inc., pp. 213-230, 2010.

[28] T. Griebel, M. Brinkmeyer and S. Böcker, EPoS : a modular software framework for phylogenetic analysis. *Bioinformatics*, 24(20) :2399–2400, 2008.

[29] J.S. Farris, Phylogenetic Analysis Under Dollo's Law. *Systematic Zoology*, 26(1) :77-88

[30] J.C. Clemente, K. Ikeo, G. Valiente and T. Gojobori, Optimized ancestral state reconstruction using sankoff parsimony. *BMC Bioinformatics*, 10 :51 (2009).

[31] S.J. Marygold, P.C. Leyland, R.L. Seal, J.L. Goodman and *et al.*,. FlyBase : improvements to the bibliography. *Nucleic Acids Res.*, 41(Database issue) :D751-D757.

# A New Framework for Computational Protein Design

## through Cost Function Network Optimization

Seydou TRAORÉ[1-3], David ALLOUCHE[4], Isabelle ANDRÉ[1-3], Simon DE GIVRY[4], George KATSIRELOS[4], Thomas SCHIEX[4*] and Sophie BARBE[1-3*]

[1] Université de Toulouse;INSA,UPS,INP; LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France
sophie.barbe@insa-toulouse.fr

[2] INRA, UMR792, Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse

[3] CNRS, UMR5504, F-31400 Toulouse, France

[4] Mathématiques et Informatique Appliquées de Toulouse, UR 875, INRA, F-31320 Castanet Tolosan, France
thomas.schiex@toulouse.inra.fr

**Keywords** Structural biology, combinatorial optimization, computational protein design.

The engineering of tailored proteins with desired properties holds great interest for applications ranging from medicine, biotechnology and synthetic biology and nanotechnologies. In recent years, structure-based computational protein design (CPD) approaches have demonstrated their potential to adequately capture fundamental aspects of molecular recognition and interactions and have already enabled the successful (re)design of several enzymes (see for example [1]).

One of the main challenges of CPD lies in the exponential size of the conformational and sequence protein space that has to be explored which rapidly grows out of reach of computational approaches. In the simplest form, the CPD problem assumes a fixed protein backbone and, for each type of amino acid considered at a given position, allows the side-chains to move only among a set of discrete and low-energy conformations, called rotamers. CPD is thus formulated as an optimization problem which consists in choosing combinations of rotamers at designable specified positions such that the fold has minimum energy (global minimum energy conformation or GMEC). This problem has been proven to be NP-hard [2]. If several meta-heuristic methods have been applied to it, there also exist methods which solve the GMEC exactly. The most usual is based on the Dead-End Elimination (DEE) theorem [3]. Such exact methods offer several advantages. First, they ensure that discrepancies between CPD predictions and experimental results come exclusively from the inadequacies of the biophysical model and not from the algorithm. Next, because provable methods can determine that the optimum is reached, they may actually stop before meta-heuristic approaches. Finally, empirical studies on solving the GMEC problem reported that the accuracy of meta-heuristic approaches tend to degrade as the problem size increases [4].

In this work, we show that a recent combinatorial optimization technique, defined in the field of "Cost Function Networks" (or Weighted Constraint satisfaction, [5]) can push CPD beyond the limit of usual tools. We propose a new design strategy which starts from a PDB structure, selects mutable and flexible residues identified on the basis of the functional and structural knowledge on the target protein, in particular the amino acid burial in the structure which is captured by the solvation radius [6] and exploits energy fields to reach a final CPD instance. A CPD instance defines an exponential space of possible sequences with associated conformations (choice of rotamer) and the ability to compute the energy of any sequence-conformation configuration using a pairwise energy matrix.

Following this methodology, we have built 35 design cases, involving free proteins, or proteins bound to a cofactor, a ligand or a protein. The studied systems have all been extracted from previously published papers about protein engineering, *in silico* protein design or protein structural studies. The size of the design cases include from 3 to 119 mutable residues and encompass spaces from $4.10^{26}$ to $2.10^{249}$.

For each of these cases, we tried to identify the GMEC using either Osprey 2.0 (a common open source CPD software [7]), or modeled the problem as a Cost Function Network and solved using toulbar2 (a dedicated open source CFN solver[1]) or reformulated it as an Integer Linear programming problem (ILP) and solved it using the commercial IBM ILOG Cplex ILP solver. Within a time-out of 100 hours per design,

[1] Toulbar2 is available at https://mulcyber.toulouse.inra.fr/projects/toulbar2.

Osprey solved 18 cases, CPLEX solved 27 cases and toulbar2 solved 30 instances. No instance unsolved by toulbar2 could be solved by other approaches. With a shorter time-out of 10 minutes, these numbers reduced to 11, 13 and 30, showing the efficiency of the CFN approach.

In practice, finding the GMEC is not always sufficient and a gap-free ensemble of solutions close to (and including) the GMEC is sought in order to design a larger library of protein mutants to be tested experimentally. Indeed, the energetic model used is only an approximation and the GMEC may not be the actual most stable configuration. Furthermore, the most stable configuration may be so stable that it loses the original function of the parental wild type protein. We therefore compared the capabilities of Osprey and toulbar2 for generating the set of all suboptimal solutions within a 2 kcal/mol threshold of the optimum. Among the set of 30 cases for which a GMEC could be previously identified, Osprey 2.0 was able to produce such an ensemble for just one of the simplest design cases (taking around 37 hours), while toulbar2 successfully produced all ensembles for the 30 cases, taking less than 7 hours in all cases.

The produced ensembles could contain up to $8.10^8$ different conformations, but never represented more than $3.10^5$ different sequences (and often much less). We more thoroughly analyzed 4 cases for which the number of sequences was below 300. Their energy was lower than the wild type model by as much as 20 kcal/mol. We expected at some core positions that mutations would favor the introduction of bulkier amino acids in order to fill up the free space. However, changes in amino acid sizes were subtle, probably because of the lack of molecular flexibility in the underlying model and the discretization generated by rotamers. As this assumption leads to ignoring the structural rearrangements that may occur in protein mutants, we further minimized the energy of the best conformation of each unique sequence in continuous space while considering flexible all amino acid backbone and side chains. Despite that the superposition of the structures before and after minimization only showed slight conformational rearrangements, the resulting decreases in energy was important (up to 60 kcal/mol). Therefore, sequences were re-ranked taking into consideration a subsequent geometry optimization (difficult to handle during CPD).

Beyond providing a design framework and computational tools to facilitate the optimization of highly combinatorial design cases, our approach also has the potential to speedup methods that integrate more flexibility as long as they reduce to the same type of optimization problems [8]. This is even more important as this considerably expands the size of the search space or may require solving a large number of GMECs.

## Acknowledgments

This work was supported by the “Agence Nationale de la Recherche”, references ANR 10-BLA-0214 and ANR-12-MONU-0015-03. We thank the Computing Center of Region Midi-Pyrénées (CALMIP, Toulouse, France) and the GenoToul Bioinformatics Platform of INRA-Toulouse for providing computing resources and support. S. Traoré was supported by a grant from the INRA and the Region Midi-Pyrénées.

## References

[1] Khare,S.D. et al. Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nat. Chem. Biol*., 8, 294–300. 2012.

[2] Pierce,N.A. and Winfree,E. (2002) Protein Design is NP-hard. Protein Engineering, 15, 779–782.

[3] Desmet,J. et al. The dead-end elimination theorem and its use in protein sidechain positioning. *Nature*, 356, 539–542. 1992.

[4] Voigt,Christopher A. et al. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *Journal of Molecular Biology*, 299, 789–803, 2000.

[5] Allouche,D. et al. Computational Protein Design as a Cost Function Network Optimization Problem. In *Proc. of CP*. 2012.

[6] Archontis,G. and Simonson,T. A residue-pairwise Generalized Born scheme suitable for protein design calculations. *J. Phys. Chem. B*, 109, 22667–22673. 2005.

[7] P. Gainza et al. OSPREY: Protein design with ensembles, flexibility, and provable algorithms. *Methods Enzymol.* 2013;523:87-107.

[8] Hallen,M.A. et al. Dead-end elimination with perturbations (DEEPer): a provable protein design algorithm with continuous sidechain and backbone flexibility. *Proteins*, 81, 18–39. 2013.

# Smoothing 3D Protein Structure Motifs Through Graph Mining and Amino-Acids Similarities ⋆

Wajdi DHIFLI[1,2], Rabie SAIDI[3] and Engelbert MEPHU NGUIFO[1,2]

[1] Clermont University, Blaise Pascal University, LIMOS, BP 10448, F-63000 Clermont-Ferrand, France
[2] CNRS, UMR 6158, LIMOS, F-63173 Aubière, France
dhifli, mephu@isima.fr
[3] European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, United Kingdom
rsaidi@ebi.ac.uk

**Abstract** *One of the most powerful techniques to study proteins is to look for recurrent fragments (also called substructures or spatial motifs), then use them as patterns to characterize the proteins under study. An emergent trend consists in parsing proteins three-dimensional (3D) structures into graphs of amino acids. Hence, the search of recurrent substructures is formulated as a process of frequent subgraph discovery where each subgraph represents a 3D-motif. In this scope, several efficient approaches for frequent 3D-motifs discovery have been proposed in the literature. However, the set of discovered 3D-motifs is too large to be efficiently analyzed and explored in any further process. In this paper, we propose a novel pattern selection approach that shrinks the large number of discovered frequent 3D-motifs by selecting the representative ones. Existing pattern selection approaches do not exploit the domain knowledge. Yet, in our approach we incorporate the evolutionary information of amino acids defined in the substitution matrices in order to select the representative 3D-motifs. We show the effectiveness of our approach on a number of real datasets. The results issued from our experiments show that our approach detects relations between patterns that current subgraph selection approaches fail to detect, and that it is able to considerably decrease the number of motifs while enhancing their interestingness.*

**Keywords** Feature selection, 3D-motifs, protein structures, representative unsubstituted patterns.

## 1 Introduction

Studying protein structures can reveal relevant structural and functional information which may not be derived from protein sequences alone. During recent years, various methods that study protein structures have been elaborated based on diverse types of descriptor such as profiles [15] and motifs [11] and others [17,10]. Yet, the exponential growth of online databases such as the Protein Data Bank [2] and SCOP [1] arises an urgent need for more accurate methods that will help to better understand the studied phenomenons such as protein evolution, functions, etc. In this scope, proteins have recently been interpreted as graphs of amino acids [7]. This representation enables the use of graph mining techniques to study protein structures. One of the powerful and current trends in graph mining is frequent subgraph discovery which aims to discover subgraphs that frequently occur in a graph dataset. The discovered subgraphs are used lately to reveal interesting information hidden in the original graphs, such as discovering pathways in metabolic networks [5], identifying residues that discriminate protein families [14], etc.

The graph isomorphism test is one of the main bottlenecks of frequent subgraph mining. Yet, many efficient and scalable algorithms have been proposed in the literature and made it feasible for instance FFSM [8], gSpan [20], etc. Unfortunately, the exponential number of discovered frequent 3D-motifs is another serious issue that still needs more attention [18], since it may hinder or even make any further analysis unfeasible due to time, resources and computational limitations. This problem becomes more serious with data of higher density such as proteins 3D structures. In fact, the issues raised from the huge number of frequent 3D-motifs are mainly due to two factors, namely *redundancy* and *significance* [14]. The redundancy is mainly caused by structural

⋆. Datasets and an implementation of the approach are freely available upon email request or at: http://fc.isima.fr/∼dhifli/unsubpatt/

and/or semantic similarity, since most discovered 3D-motifs differ slightly in structure and may infer similar or even the same meaning. Moreover, the significance of the frequent 3D-motifs is only related to frequency. This yields an urgent need for efficient selection of relevant patterns among the frequent ones.

Statistical pattern selection methods have been widely used to resolve the dimensionality problem when the number of discovered patterns is too large. However, current methods are too generic and do not consider the specificity of the domain and the used data. We believe that the incorporation of the domain knowledge will allow creating approaches that best fit the considered data. In proteomics, a protein is composed by the folding of a set of amino acids. During evolution, amino acids can substitute each other. These substitutions were quantified by biologists in the form of substitution matrices. In this paper, we propose a novel selection approach which selects a subset of representative 3D-motifs (in the form of subgraphs) among the frequent ones. We term them *unsubstituted patterns*. In our selection, we exploit a specific domain knowledge which is the substitution between amino acids represented as nodes. Though, the main contribution of this work is to mine a representative summary of the set of frequent 3D-motifs by incorporating a prior domain knowledge which is the ability of substitution between amino acids. It is worth mentioning that the proposed approach can be used on any biological data whenever it is possible to define a matrix quantifying the substitutions between the nodes labels. Our approach also handles other types of structures such as trees and paths (sequences) and is unsupervised thus it can help in various mining tasks such as classification and clustering.

The remainder of the paper is organized as follows. Section 2 discusses recent works dealing with 3D-motifs selection that were applied on protein structures. Section 3 presents the background of our work, the preliminaries and the main algorithm. Then, Section 4 describes the characteristics of the used data and the experimental settings. Section 5 presents and discuss the obtained experimental results. In the rest of the paper, we use the following terms interchangeably : 3D motif, subgraph, pattern.

## 2 Related Works

Recently, several subgraph selection approaches have been proposed and applied on protein structures. On the whole, they transform the protein 3D structures on graphs, then they try to mine the most relevant 3D-motifs in the form of subgraphs. ORIGAMI [3] is an approach for both subgraph discovery and selection. It straightforwardly selects the $\alpha$-orthogonal (non-redundant) and $\beta$-representative patterns among a sample of maximal frequent subgraphs. The LEAP algorithm proposed in [19] tries to locate patterns that individually have high discrimination power, using an objective function score that measures each pattern's significance. Another approach termed gPLS [13] attempts to select a set of informative subgraphs in order to rapidly build a classifier. gPLS uses the mathematical concept of partial least squares regression to create latent variables allowing a better prediction. COM [9] is another subgraph selection approach which follows a process of pattern mining and classifier learning. It mines co-occurrence rules, then uses them to assemble weak features in order to generate strong ones. In [14], authors proposed a feature selection approach termed CORK which selects the subgraphs that are most discriminative for classification using a submodular quality function. In [6], authors designed LPGBCMP, a general model which selects clustered features by considering the structure relationship between subgraph patterns in the functional space. The selected subgraphs are used as base learners to obtain high quality classification models. To the best of our knowledge, in all current subgraph selection approaches, the selection is usually based on structural similarity (approximate structural isomorphism, ...) and/or statistical measures (discrimination, correlation, ...). Yet, the *prior* domain knowledge are often ignored. However, exploiting them may help building more accurate approaches that best fit the studied data.

## 3 Mining Representative Unsubstituted Patterns

### 3.1 Background

The main idea of our approach is based on node substitution. Since in a graph representing a protein structure, amino acids are presented as nodes. Though, using a substitution matrix, it would be possible to quantify the substitution between two given subgraphs. Starting from this idea, we define a similarity function that measures the distance between a given pair of 3D-motifs. Then, we preserve only one 3D-motif from each

substitutable pair with respect to a similarity threshold, such that the preserved 3D-motifs represent the set of representative unsubstituted patterns. An overview of the proposed approach is illustrated in Fig. 1 and a more detailed description is given in the following sections. The substitution between amino acids was also used

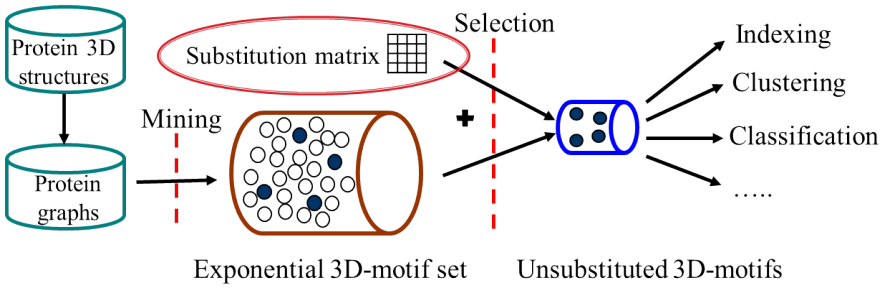



**Figure 1.** Unsubstituted 3D-motif selection framework.

in the literature but for sequential motifs extraction from protein sequences in [12], where authors proposed a feature extraction approach termed *DDSM* for protein sequence classification. DDSM generates every frequent subsequence substituting another one, i.e, it eliminates any sequential motif substituted by another one and which itself does not substitute any other motif. We believe that their approach does not guarantee an optimal summarization since its output may still contain substitutable patterns. In addition, they do consider negative substitution scores as impossible substitutions which is biologically not true since negative scores are only expressing the less likely substitutions. Moreover, DDSM is limited to protein primary structures (sequences) and does not handle the tertiary structures. Our approach overcomes these shortcomings, since it handles both protein sequences (seen as line graphs) and 3D structures. In addition, it considers both the positive and negative substitution scores. Moreover, our approach generates a set of representative unsubstituted patterns ensuring an optimal summarization of the initial set. Besides, it is unsupervised and can be exploited in classification as well as in other analysis and mining tasks unlike DDSM which is dedicated to classification.

### 3.2 Preliminaries

In this section we present the preliminaries and the formal problem statement. Let $\mathcal{G}$ be a dataset of protein structures represented as graphs. Each graph $G = (V, E, L)$ of $\mathcal{G}$ is given as a collection of nodes (amino acid residues) $V$ and edges (interactions) $E$. The nodes of $V$ are labeled within an alphabet $L$ (amino acids types). We denote by $|V|$ the number of nodes (the graph order) and by $|E|$ the number of edges (the graph size). Let also $\Omega$ be the set of frequent subgraphs (3D-motifs) extracted from $\mathcal{G}$, also referred here as $patterns$.

DEFINITION 3.1. *(Substitution matrix) Given an alphabet $L$, a substitution matrix $\mathcal{A}$ over $L$ is the function:*

$$\mathcal{A} : \left| \begin{array}{ccc} L^2 & \longrightarrow & [\bot, \top] \subset \mathbb{R} \\ (l, l') & \longrightarrow & x \end{array} \right. \quad (1)$$

*The higher the value of $x$ is, the more likely is the substitution of $l'$ by $l$. If $x = \bot$ then the substitution is impossible, and if $x = \top$ then it is certain. The values $\bot$ and $\top$ are optional and user-specified. They may appear or not in $\mathcal{A}$. The scores in $\mathcal{A}$ should respect the following two properties:*

1. $\forall\, l \in L, \exists\, l' \in L \mid \mathcal{A}(l, l') \neq \bot$,
2. $\forall\, l \in L$, *if* $\exists\, l' \in L \mid \mathcal{A}(l, l') = \top$ *then* $\forall\, l'' \in L \backslash \{l, l'\}$, $\mathcal{A}(l, l'') = \bot$ *and* $\mathcal{A}(l', l'') = \bot$.

*In protein's substitution matrices, both positive and negative values represent possible substitutions. However, positive scores are given to the more likely substitutions while negative scores are given to the less likely substitutions. In order to give more magnitude to higher values of $x$, we define the matrix $\mathcal{M}$ over $L$ such that $\forall\, l$ and $l' \in L$:*

$$\mathcal{M}(l, l') = \mathrm{e}^{\mathcal{A}(l, l')} \quad (2)$$

DEFINITION 3.2. *(Structural isomorphism) Two patterns $P = (V_P, E_P, L)$ and $P' = (V_{P'}, E_{P'}, L)$ are said to be structurally isomorphic (having the same shape), we note $shape(P, P') = true$, iff:*

- *$P$ and $P'$ have the same order, $i.e.$, $|V_P| = |V_{P'}|$,*
- *$P$ and $P'$ have the same size, $i.e.$, $|E_P| = |E_{P'}|$,*
- *$\exists$ a bijective function $f : V_P \rightarrow V_{P'} \mid \forall u, v \in V_P$ if $(u, v) \in E_P$ then $(f(u), f(v)) \in E_{P'}$ and inversely.*

*It is worth mentioning that in this definition, we consider only the isomorphism on structure and we ignore the labels while the standard graph isomorphism considers them both.*

DEFINITION 3.3. *(Elementary conservation probability) Given a node $v$ of a label $l \in L$, the elementary conservation probability, $M_{el}(v)$, measures the possibility that $v$* ***does not*** *mutate to any other node depending on its label $l$.*

$$M_{el}(v) = \begin{cases} 0, \text{ if } \mathcal{M}(l,l) = \mathrm{e}^{\perp} \\ 1, \text{ if } \mathcal{M}(l,l) = \mathrm{e}^{\top} \\ \frac{\mathcal{M}(l,l)}{\sum_{i=1}^{|L|} \mathcal{M}(l,l_i)}, \text{otherwise} \end{cases} \tag{3}$$

*Obviously, if the substitution score in $\mathcal{M}$ between $l$ and itself is $\perp$ then $l$ will certainly mutate to another label $l'$ and $M_{el}(v) = 0$. Respectively, if this substitution score is $\top$ then it is certain that $v$ will not mutate, so $M_{el}(v) = 1$. Otherwise, we divide the score that $l$ mutates to itself by the sum of all the possible mutations.*

DEFINITION 3.4. *(Pattern mutation probability) Given a pattern $P = (V_P, E_P, L) \in \Omega$, the pattern mutation probability, $M_{patt}(P)$, measures the possibility that $P$ mutates to any other pattern having the same order.*

$$M_{patt}(P) = 1 - \prod_{i=1}^{|V_P|} M_{el}(V_P[i]) \tag{4}$$

*where $\prod_{i=1}^{|V_P|} M_{el}(V_P[i])$ represents the probability that the pattern $P$ does not mutate to any other pattern $i.e.$ $P$ stays itself.*

DEFINITION 3.5. *(Elementary substitution score) Given two nodes $v$ and $v'$ having correspondingly the labels $l, l' \in L$, the elementary substitution score, $S_{el}(v, v')$, measures how high $v$ can substitute $v'$.*

$$S_{el}(v, v') = \frac{\mathcal{M}(l, l')}{\mathcal{M}(l, l)} \tag{5}$$

DEFINITION 3.6. *(Pattern substitution score) Given two structurally isomorphic patterns $P$ and $P'$, we denote by $S_{patt}(P, P')$ the substitution score of $P'$ by $P$. In other words, it measures the possibility that $P$ mutates to $P'$ by computing the normalized sum of the elementary substitution scores of the nodes of $P$. Formally:*

$$S_{patt}(P, P') = \frac{\sum_{i=1}^{|V_P|} S_{el}(V_P[i], V_{P'}[i])}{\mid V_P \mid} \tag{6}$$

DEFINITION 3.7. *(Pattern substitution) A pattern $P$ substitutes $P'$, we note $subst(P, P', \tau) = true$, iff:*

1. *$P$ and $P'$ are structurally isomorphic ($shape(P, P') = true$),*
2. *$S_{patt}(P, P') \geq \tau$, where $\tau$ is a user-specified threshold such that $0\% \leq \tau \leq 100\%$.*

DEFINITION 3.8. *(Unsubstituted pattern) Given a threshold $\tau$ and $\Omega^* \in \Omega$, a pattern $P^* \in \Omega^*$ is said to be unsubstituted iff: $\nexists P \in \Omega^* \mid M_{patt}(P) > M_{patt}(P^*)$ and $subst(P, P^*, \tau) = true$.*

PROPOSITION 3.9. *Given a pattern $P = (V_p, E_p, L) \in \Omega$, if $M_{patt}(P) = 0$ then $P$ is an unsubstituted pattern.*

*Proof.* The proof can be simply deduced from Definitions 3.3 and 3.4. □

DEFINITION 3.10. *(Merge support) Given two patterns $P$ and $P'$, if $P$ substitutes $P'$ then $P$ will represent $P'$ in the list of graphs where $P'$ occurs. Formally:*

$$\forall P, P' \in \Omega, D_P = D_P \cup D_{P'} \mid subst(P, P', \tau) = true \tag{7}$$

*where $D_P$ and $D_{P'}$ are correspondingly the occurrence list of $P$ and that of $P'$.*

### 3.3 Algorithm

Given a set of patterns $\Omega$ and a substitution matrix $\mathcal{M}$, we propose UNSUBPATT a pattern selection algorithm which enables detecting the set of unsubstituted patterns $\Omega^*$ within $\Omega$. Based on our similarity concept, all the patterns in $\Omega^*$ are dissimilar, since it does not contain any pair of patterns that are substitutable. This represents a reliable summarization of $\Omega$.

The general process of the algorithm is described as follows: first, $\Omega$ is divided into subsets of patterns having the same number of nodes and edges. Then, each subset is sorted in a descending order by mutation probability $M_{patt}$. Each subset is browsed starting from the pattern having the highest $M_{patt}$. For each pattern, we remove all the patterns it substitutes and we merge their occurrence lists such that the preserved pattern will represent all the removed ones wherever they occurs. The remaining patterns represent the set of unsubstituted ones. Our algorithm uses Proposition 3.9 to avoid unnecessary computation related to patterns with $M_{patt} = 0$. They are directly considered as unsubstituted patterns, since they can not mutate to any other one.

PROPERTY 3.11. *Let $\Omega$ be a set of patterns and $\Omega^*$ its subset of unsubstituted patterns based on a substitution matrix $\mathcal{M}$ and a threshold $\tau$, i.e.,* UNSUBPATT $(\Omega, \mathcal{M}, \tau, (\perp, \top)) = \Omega^*$. *Then :*

$$\text{UNSUBPATT}(\Omega^*, \mathcal{M}, \tau, (\perp, \top)) = \Omega^* \tag{8}$$

## 4 Experiments

### 4.1 Datasets

In order to experimentally evaluate our approach, we use four datasets of protein 3D structures, which also have been used in [19] then [6]. Each dataset consists of two classes: positive and negative. Positive samples are proteins selected from a considered protein family whereas negative samples are proteins randomly gathered from the Protein Data Bank [2]. Each protein is parsed into a graph of amino acids. Each node represents an amino acid residue and is labeled with its amino acid type. Two nodes $u$ and $v$ are linked by an edge $e(u, v) = 1$ if the euclidean distance between their two $C_\alpha$ atoms $\Delta(C_\alpha(u), C_\alpha(v))$ is below a threshold distance $\delta$. In the literature, many methods use this definition with usually $\delta \geq 7$Å on the argument that $C_\alpha$ atoms define the overall shape of the protein conformation. In our experiments, we use $\delta = 7$Å.

Table 1 summarizes the characteristics of each dataset. SCOP ID, Avg.|V| and Avg.|E| correspond respectively to the identifier of the protein family in SCOP database [1], the average number of nodes and the average number of edges in each dataset.

| Dataset | SCOP ID | Family name | Pos. | Neg. | Avg.\|V\| | Avg.\|E\| |
|---|---|---|---|---|---|---|
| DS1 | 52592 | G proteins | 33 | 33 | 246 | 971 |
| DS2 | 48942 | C1-set domains | 38 | 38 | 238 | 928 |
| DS3 | 56437 | C-type lectin domains | 38 | 38 | 185 | 719 |
| DS4 | 88854 | Protein kinases, catalytic subunit | 41 | 41 | 275 | 1077 |

**Table 1.** Characteristics of the experimental datasets

DS1 contains proteins from the G protein family, also known as guanine nucleotide-binding proteins, which are involved in transmitting chemical signals originating from outside a cell into the inside of it. The C1-set domains composing DS2 resemble the antibody constant domains. They are mostly found in molecules involved in the immune system, in the major histocompatibility complex class I and II complex molecules, and in various T-cell receptors. In DS3, the C-type (Calcium-dependent) lectins are a family of lectins which share structural homology in their high-affinity carbohydrate-recognition domains. In fact, this family involves groups of proteins playing divers functions including cell-cell adhesion, immune response to pathogens and apoptosis. Protein kinases, catalytic subunit composing DS4 are mainly proteins that modifies other proteins by chemically adding phosphate groups to them. This usually results in a functional change of the target protein.

| Dataset | $\mid \Omega \mid$ | $\mid \Omega^* \mid$ | Selection rate (%) |
|---|---|---|---|
| DS1 | 799094 | 7291 | 0.91 |
| DS2 | 258371 | 15898 | 6.15 |
| DS3 | 114792 | 14713 | 12.82 |
| DS4 | 1073393 | 9958 | 0.93 |

**Table 2.** Number of frequent 3D-motifs ($\Omega$), representative unsubstituted patterns ($\Omega^*$) and the selection rate

| Dataset | Accuracy | | Sensitivity | | AUC | |
|---|---|---|---|---|---|---|
| | FSg | UnSubPatt | FSg | UnSubPatt | FSg | UnSubPatt |
| DS1 | 0.62 | 0.78 | 0.70 | 0.90 | 0.64 | 0.78 |
| DS2 | 0.80 | 0.90 | 0.74 | 0.86 | 0.79 | 0.89 |
| DS3 | 0.86 | 0.94 | 0.86 | 0.89 | 0.86 | 0.94 |
| DS4 | 0.79 | 0.98 | 0.70 | 0.98 | 0.76 | 0.94 |

**Table 3.** Accuracy, sensitivity and AUC of the classification of each dataset using NB coupled with frequent 3D-motifs (FSg) then representative unsubstituted patterns (UnSubPatt)

### 4.2 Protocol and Settings

Generally, in a pattern selection approach two aspects are emphasized, namely the number of selected patterns and their interestingness. In order to evaluate our approach, we first use the state-of-the-art method of frequent subgraph discovery gSpan [20] to find the 3D-motifs in each protein dataset with a minimum frequency threshold of 30%. Then, we use UNSUBPATT to select the representative unsubstituted patterns among them with a minimum substitution threshold $\tau$=30%. We use $Blosum62$ [4] as the substitution matrix because it performs well on detecting the majority of weak protein similarities. It is worth mentioning that the choice of 30% as minimum frequency threshold for the frequent 3D-motifs extraction is to make the experimental evaluation feasible due to time and computational limitations.

In order to evaluate the number of selected patterns, we define the selection rate as the rate of the number of representative unsubstituted patterns from the initial set of frequent 3D-motifs. To evaluate the interestingness of the set of selected patterns, we use them as features for classification. We perform a 5-fold cross-validation classification (5 runs) on each protein-structure dataset. Each protein is encode into a binary vector, denoting by "1" or "0" the presence or the absence of the feature in the considered protein. For classification, we use one of the most known classifier, namely the naïve bayes (NB) classifier, due to its simplicity and fast prediction and that its classification technique is based on a global and conditional evaluation of the input features. NB is used with the default parameters from the workbench Weka [16].

## 5 Results and Discussion

### 5.1 Empirical Results

Here, we show the results of our experiments in terms of number of patterns and classification results. As mentioned before, we use gSpan to extract the frequent 3D-motifs from each dataset with frequency $\geq$ 30%. Then, we use UNSUBPATT to select the representative unsubstituted patterns among them with a substitution threshold $\tau$=30% and using Blosum62 as substitution matrix. At last, we evaluate the quality of each subset by performing a 5-fold cross-validation classification (5 runs) using the naïve bayes classifier. The obtained results are reported in Table 2 and Table 3. The high number of discovered frequent 3D-motifs is due to their combinatorial nature (as discussed in the introduction). The results reported in Table 2 show that our approach decreases considerably the number of 3D-motifs and selects only a small subset of them. The selection rate shows that the number of unsubstituted patterns $\mid \Omega^* \mid$ does not exceed 13% of the initial set of frequent 3D-motifs $\mid \Omega \mid$ in the worst case with DS3 and even reaches less than 1% with DS1 and DS4. This proves that exploiting the domain knowledge, which in this case is the substitution matrix, enables emphasizing many relations between patterns that are possibly ignored by current pattern selection approaches.

| Dataset | FSg | | UnSubPatt_Blosum80 | | UnSubPatt_Pam250 | |
|---|---|---|---|---|---|---|
| | #patterns | Accuracy | #patterns | Accuracy | #patterns | Accuracy |
| DS1 | 799094 | 0.62 | 7328 | 0.67 | 6137 | 0.68 |
| DS2 | 258371 | 0.80 | 15930 | 0.87 | 15293 | 0.87 |
| DS3 | 114793 | 0.86 | 14792 | 0.91 | 14363 | 0.93 |
| DS4 | 1073393 | 0.79 | 10417 | 0.90 | 9148 | 0.90 |

**Table 4.** Number of patterns (#patterns) and accuracy of classification of each dataset using NB with all frequent 3D-motifs (FSg) then unsubstituted patterns using Blosum80 (UnSubPatt_Blosum80) and Pam250 (UnSubPatt_Pam250)

The classification results reported in Table 3 help to evaluate the quality of the selected patterns. Indeed, this will demonstrate if the unsubstituted patterns were arbitrarily selected or they are really representative. Table 3 shows that the classification accuracy significantly increases with all datasets. We notice a huge leap in accuracy especially with DS1 and DS4 with a gain of more than 17% and reaching almost full accuracy with DS4. To better understand the accuracy results, we also report the average sensitivity and AUC values for all cases. We also notice an enhancement of performance with the mentioned quality metrics. This supports the reliability of our selection approach.

### 5.2 Results Using Other Substitution Matrices

Besides Blosum62, biologists also defined other substitution matrices describing the likelihood that two amino acid types would mutate to each other in evolutionary time. In order to study the effect of using other substitution matrices on the performance of UNSUBPATT, we perform the same experiments using two other substitution matrices, namely $Blosum80$ and $Pam250$. We compare the obtained results in terms of number of patterns and classification accuracy with those obtained using the set of frequent 3D-motifs. The results are reported in Table 4. A high selection rate accompanied with a clear enhancement of the classification accuracy is noticed using UNSUBPATT with all the substitution matrices compared to the results using the whole set of frequent 3D-motifs. It is clearly noticed that even using different substitution matrices, UNSUBPATT shows relatively similar behavior and is able to select a small yet relevant subset of patterns. It is also worth mentioning that for all the datasets, the best classification accuracy is obtained using Blosum62 (see Table 2 and Table 3) and the best selection rate is achieved using Pam250. This is simply due to how distant proteins within the same dataset are, since each substitution matrix was constructed to implicitly express a particular theory of evolution. Though, choosing the appropriate substitution matrix can influence the outcome of the analysis.

### 5.3 Impact of Substitution Threshold

In the shown experiments, we used 30% as minimum substitution threshold for UNSUBPATT. In this section, we study the impact of variation of the substitution threshold on both the number of selected patterns and the classification results. To do so, we perform the same experiments while varying the substitution threshold from 0% to 90% with a step-size of 10. In order to check if the enhancements of the obtained results are due to our selected features or to the classifier, we use two other well-known classifiers namely the support vector machine (SVM) and decision tree (C4.5) besides the naïve bayes (NB) classifier. We use the same protocol and settings of the previous experiments. Fig. 2 presents the selection rate for all substitution thresholds and Fig. 3, Fig. 4 and Fig. 5 illustrate the classification accuracies obtained respectively using NB, SVM and C4.5 with each dataset. The classification accuracy of the initial set of frequent 3D-motifs (gSpan, the line in red) is considered as a standard value for comparison. Thus, the accuracy values of UNSUBPATT (in blue) that are equal or above the standard values are considered as gains, and those below them are considered as losses.

In Fig. 2, we notice that UNSUBPATT reduces considerably the number of patterns especially with lower substitution thresholds. In fact, the number of representative unsubstituted patterns does not exceed 30% for all substitution thresholds below 70% and even reaches less then 1% in some cases. This important reduction in the number of patterns comes with a notable enhancement of the classification accuracies. This fact is illustrated in Fig. 3, Fig. 4 and Fig. 5 which show that the unsubstituted patterns allow better classification performance compared to the original set of frequent 3D-motifs. UNSUBPATT scores very well with the three used classifiers

and even reaches full accuracy in some cases. This confirms our assumptions and shows that our selection is reliable and contributes to the enhancement of the accuracy. However, we believe that NB allows the most reliable evaluation because it performs a classification based on a global and conditional evaluation of features, unlike SVM which performs itself another attribute selection to select the support vectors and unlike C4.5 which performs an attribute by attribute evaluation.

Using a substitution threshold of 0% enables UNSUBPATT to select only one representative pattern from each group of structurally isomorphic 3D-motifs. Based on the experimental results, we believe that considering only these patterns allows a fast selection and performs very well in structural classification tasks.

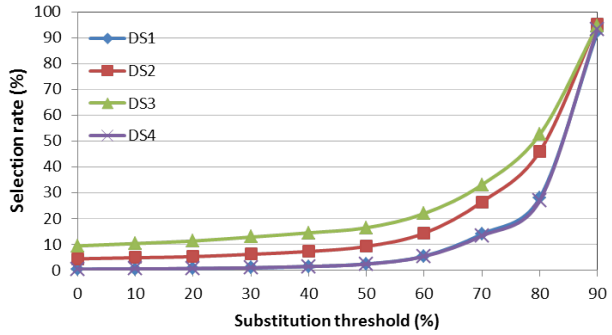



**Figure 2.** Rate of unsubstituted patterns from $\Omega$ depending on the substitution threshold ($\tau$).

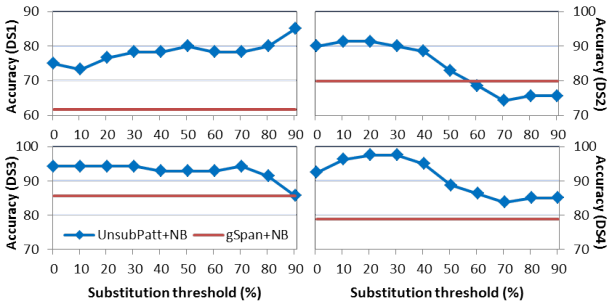



**Figure 3.** Classification accuracy by NB.

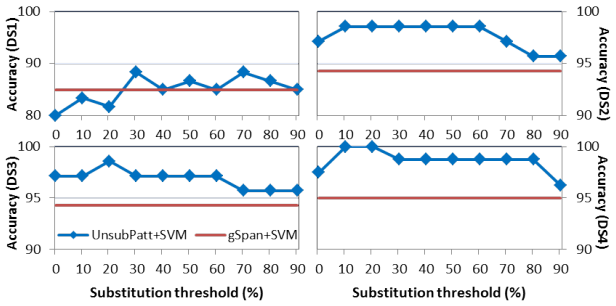



**Figure 4.** Classification accuracy by SVM.

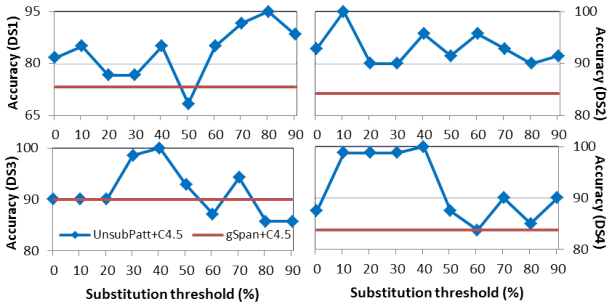



**Figure 5.** Classification accuracy by C4.5.

## 5.4 Smoothing the Distribution of Patterns

In this section, we study the distribution of patterns based on their size (number of edges). We try to check which sizes of patterns are more concerned by the selection. The Fig. 6 and Fig. 7 draw the distribution of patterns for the original set of frequent 3D-motifs and for the final set of representative unsubstituted ones with all the substitution thresholds using Blosum62. The downward tendency of UNSUBPATT using lower substitution thresholds and with respect to the original set of frequent 3D-motifs is very clear. In fact, an effect of smoothing is clearly noticed over the whole sets, since UNSUBPATT leans towards cutting off the peaks and flattening the curves with lower substitution thresholds. Another interesting observation is that the curves are flattened in the regions of small 3D-motifs as well as in the regions of big and dense ones. This demonstrates the effectiveness of UNSUBPATT with both small and big 3D-motifs.

## 5.5 Comparison with Other Approaches

To objectively evaluate our approach, we compare it with current trends in pattern selection. In Fig. 8, we report the classification accuracy using the representative unsubstituted 3D-motifs of UNSUBPATT besides those using patterns of the other considered approaches from the literature namely LEAP[19], gPLS[13], COM[9]

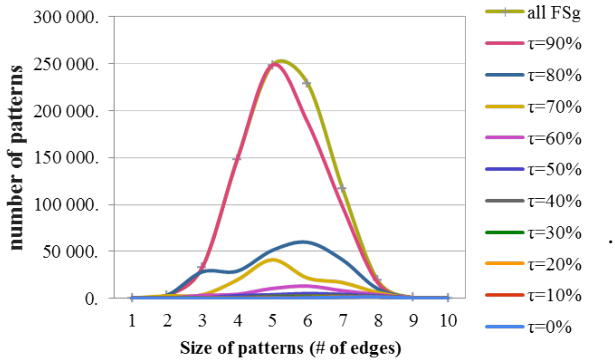



**Figure 6.** Distribution of patterns of DS1 for all the frequent 3D-motifs and for the representative unsubstituted ones with all the substitution thresholds

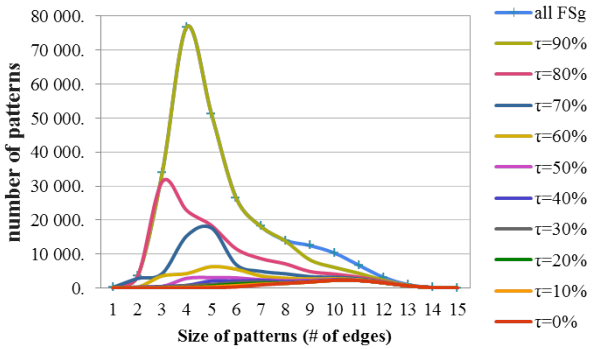



**Figure 7.** Distribution of patterns of DS2 for all the frequent 3D-motifs and for the representative unsubstituted ones with all the substitution thresholds

and LPGBCMP[6] (reported and explained in the section 2). For UNSUBPATT, we report the results obtained using the substitution matrix Blosum62, a minimum substitution threshold $\tau = 30\%$ and SVM for classification. For LEAP+SVM, LEAP is used iteratively to discover a set of discriminative subgraphs with a leap length=0.1. The discovered subgraphs are consider as features to train SVM with a 5-fold cross validation. COM is used with $t_p = 30\%$ and $t_n = 0\%$. For gPLS, the frequency threshold is set to 30% and the best accuracies are reported for all the datasets among all the parameters combinations from m = 2, 4, 8, 16 and k = 2, 4, 8, 16, where m is the number of iterations and k is the number of patterns obtained per search. For LPGBCMP, threshold values of $max_var = 1$ and $\delta = 0.25$ were respectively used for feature consistency map building and for overlapping. The obtained results are reported in the Fig. 8.

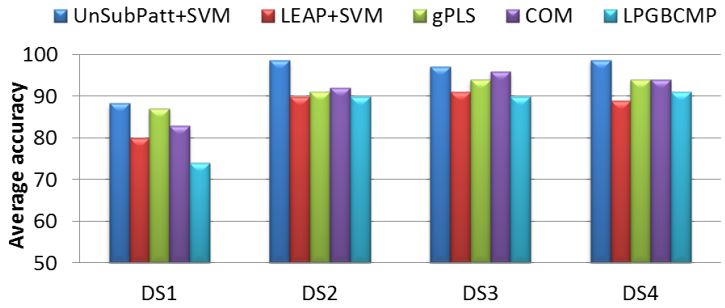



**Figure 8.** Classification accuracy comparison with other pattern selection approaches.

The classification results displayed in Fig. 8 show that UNSUBPATT allows a better classification than all the other pattern selection methods in the four cases. Considering only these results does not allow to confirm that UNSUBPATT would always outperform the considered methods. However, this proves that UNSUBPATT represents a very competitive and promising approach and is able to detects relations among patterns that the current pattern selection approaches fails to detect. It is also worth noting that these approaches are supervised and dedicated to classification unlike UNSUBPATT which is an unsupervised approach. This allows it to be used in classification as well as in other mining tasks such as clustering and indexing.

## 6 Conclusion

In this paper, we proposed a novel selection approach for mining a representative summary of a set of frequent 3D-motifs. Unlike current methods that are based on the relations between patterns in the transaction space, our approach considers the distance between patterns in the pattern space. The proposed approach exploits a specific domain knowledge, in the form of a substitution matrix, to select a subset of representative unsubstituted patterns from a given set of frequent 3D-motifs. The results issued from our analysis revealed that incorporating the domain knowledge allows our approach to detect relations between patterns that current

pattern selection approaches fail to detect. It also allows to reduce considerably the size of the initial set of 3D-motifs to obtain a more interesting and representative one enabling easier and more efficient further explorations. It is also worth mentioning that our approach can also be used on protein sequences (seen as line graphs) and is not limited to classification, but can help in other motif-based analysis.

A promising future direction is to consider also the insertions and deletions over the nodes and edges and thus to consider the substitution over patterns with different sizes. Although this increases exponentially the complexity and the difficulty of the selection, it is closer to the real world substitution phenomenon. Since the proposed approach is a filter approach, it would be also interesting to embed the selection within the extraction process in order to directly mine the representative patterns from data.

## References


[1] A. Andreeva, D. Howorth, 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: new developments. *Nucleic Acids Research*, 36(1):D419–D425, 2008.

[2] H. M. Berman, J. D. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[3] V. Chaoji, M. A. Hasan, S. Salem, J. Besson, and M. J. Zaki. Origami: A novel and effective approach for mining representative orthogonal graph patterns. *Statistical Analysis and Data Mining*, 1(2):67–84, 2008.

[4] S. R. Eddy. Where did the blosum62 alignment score matrix come from? *Nature Biotechnology*, pages 1035–1036, 2004.

[5] K. Faust, P. Dupont, J. Callut, and J. van Helden. Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics*, 26(9):1211–1218, 2010.

[6] H. Fei and J. Huan. Boosting with structure information in the functional space: an application to graph classification. In *ACM knowledge discovery and data mining conference (KDD)*, pages 643–652, 2010.

[7] J. Huan, W. Wang, D. B, J. Snoeyink, J. Prins, and A. Tropsha. Mining spatial motifs from protein structure graphs. In *International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 308–315, 2004.

[8] J. Huan, W. Wang, and J. Prins. Efficient mining of frequent subgraphs in the presence of isomorphism. In *IEEE International Conference on Data Mining (ICDM)*, pages 549–552, 2003.

[9] N. Jin, C. Young, and W. Wang. Graph classification based on pattern co-occurrence. In *ACM International Conference on Information and Knowledge Management*, pages 573–582, 2009.

[10] J. Maupetit, P. Derreumaux, and P. Tufféry. A fast method for large-scale *de novo* peptide and miniprotein structure prediction. *Journal of Computational Chemistry*, 31(4):726–738, 2010.

[11] L. Regad, A. Saladin, J. Maupetit, C. Geneix, and A.-C. Camproux. Dissecting protein loops with a statistical scalpel suggests a functional implication of some structural motifs. *BMC Bioinformatics*, 12:247, 2011.

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

[13] H. Saigo, N. Krämer, and K. Tsuda. Partial least squares regression for graph mining. In *ACM knowledge discovery and data mining conference (KDD)*, pages 578–586, 2008.

[14] M. Thoma, H. Cheng, A. Gretton, J. Han, H.-P. Kriegel, A. Smola, L. Song, P. S. Yu, X. Yan, and K. M. Borgwardt. Discriminative frequent subgraph mining with optimality guarantees. *Statistical Analysis and Data Mining*, 3(5):302–318, 2010.

[15] N. von Öhsen, I. Sommer, R. Zimmer, and T. Lengauer. Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics*, 20(14):2228–2235, 2004.

[16] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[17] I. Wohlers, N. Malod-Dognin, R. Andonov, and G. W. Klau. Csa: comprehensive comparison of pairwise protein structure alignments. *Nucleic Acids Research*, 40:303–309, 2012.

[18] A. Woznica, P. Nguyen, and A. Kalousis. Model mining for robust feature selection. In *ACM knowledge discovery and data mining conference (KDD)*, pages 913–921, 2012.

[19] X. Yan, H. Cheng, J. Han, and P. S. Yu. Mining significant graph patterns by leap search. *ACM SIGMOD international conference on Management of data*, pages 433–444, 2008.

[20] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. *Order A Journal On The Theory Of Ordered Sets And Its Applications*, 02:721–724, 2002.

# SAXS Merge: an automated statistical tool to merge SAXS profiles

Yannick SPILL[1;2;3], Seung Joong KIM[3], Dina SCHNEIDMAN-DUHOVNY[3], Daniel RUSSEL[3], Ben WEBB[3], Andrej SALI[3] and Michael NILGES[1]

[1] Unité de Bioinformatique Structurale, Institut Pasteur, 25 rue du Docteur Roux 75015 Paris, France

[2] Université Paris Diderot – Paris 7, Paris Rive Gauche, 5 rue Thomas Mann, 75013 Paris, France

[3] Department of Bioengineering and Therapeutic Sciences, California Institute for Quantitative Biosciences at UCSF, Department of Pharmaceutical Chemistry, University of California, San Francisco, UCSF MC 2552, Byers Hall, 1700 4th Street, Suite 503B, San Francisco, CA 94158, USA

nilges@pasteur.fr

**Keywords:** SAXS, Gaussian process, Webserver, Data processing

Small-Angle X-ray Scattering (SAXS) is informative about structures of biomolecules in solution. High-throughput data collection requires robust, accurate and automated tools for data processing. High-quality SAXS profiles are usually obtained by manual merging of scattering profiles from different concentrations and exposure times. Here, we present SAXS Merge, a fully automated statistical tool for merging SAXS profiles. The required input consists only of the buffer-subtracted profile files in a three-column format, and their order. The tool was successfully validated on a benchmark of 16 SAXS datasets and provides a substantial improvement to existing manual procedures. SAXS Merge is available in the Mobyle workflow interface at http://mobyle.pasteur.fr or as a graphical interface at http://salilab.org/saxsmerge.

# Session 3B : Environnements de recherche en biologie

# Enrichissement s®mantique de vues RDF D2RQ dans le but d'automatiser l'int®gration de bases de donn®es relationnelles distribu®es

Julien WOLLBRETT[1], Pierre LARMANDE[2] and Manuel RUIZ[1]

[1] CIRAD, UMR AGAP, F-34398 Montpellier, France
{julien.wollbrett, manuel.ruiz}@cirad.fr

[2] IRD, UMR DIADE, Montpellier, France
[pierre.larmande]@ird.fr

**Abstract** *Semantic Web standards promote the integration of distributed relational databases. A mapping between relational databases and ontologies is then necessary. In this article we will present how we used a tool that maps relational databases and ontologies, in a novel way, in order to automate query creation for distributed data sources.*

**Keywords** *Data integration, Semantic Web standards, metadata, shortest path, D2RQ, BioSemantic*

## 1 Introduction

Les Bases de données relationnelles (BDR) biologiques ont la particularité d'être très nombreuses, hétérogènes, et distribuées dans différents laboratoires. Face à ces particularités, la problématique d'intégration de données occupe une place importante dans la communauté biologique. Une grande variété de techniques ont été utilisées au cours des 15 dernières années résumé dans l'article de Goble et *al.* [1]. Le point commun de l'intégration de données et des technologies du Web Sémantique est de dépasser l'hétérogénéité sémantique de sources de données interconnectées. Le Web Sémantique facilite la représentation de la sémantique des données et peut ainsi être utilisé pour faciliter l'interopérabilité ou l'intégration de données [2]. L'intégration de BDR en utilisant les standards du Web Sémantique est confrontée à une problématique de mise en correspondance (mapping) entre des schémas de BDR et une ou plusieurs ontologies. Dans notre approche nous avons décidé de détourner l'utilisation d'un outil de mapping entre schémas de BDR et ontologies nommé D2RQ pour, en plus d'homogénéiser des schémas hétérogènes, automatiser la création de requêtes sur des BDR distribuées. Pour cela nous avons enrichi sémantiquement et de manière automatique la vue du schéma de BDR créée par D2RQ. La vue résultante pourra être utilisée, par un algorithme de recherche de plus court chemin [3] prenant en compte les particularités des schémas relationnels, pour créer une requête pertinente. Cet article est focalisé sur l'enrichissement sémantique des vues du schéma de BDR. Nous y détaillons le raisonnement nous ayant conduit à choisir les annotations sémantiques à ajouter à la vue RDF. Nous présentons également la manière de détecter les informations souhaitées dans le schéma de la BDR et la manière de les annoter dans la vue RDF.

## 2 Mapping entre BDR et ontologies

Le mapping entre BDR et ontologies permet de répondre à plusieurs motivations. L'accès aux données basées sur une ontologie est une de ces motivations [4]. Cette approche suppose que le lien entre une ontologie

et une base de données agit comme une couche intermédiaire entre l'utilisateur et les données stockées. Cette approche peut être vue comme un adaptateur dans un système d'intégration de type médiation car il cache les détails de la source en transformant des requêtes expressives sur un domaine d'intérêt en requêtes interrogeant la source de données d'origine. L'avantage principal de cette approche est la possibilité d'interroger une BD sans avoir à exporter son contenu. Plusieurs outils existants permettent de répondre à ces besoins. Parmi ces outils, seuls deux sont basés sur des standards, que ce soit pour la vue du schéma de la BD (RDF, XML), ou pour le langage de requête (SPARQL), et permettent de mapper une base de données avec plus d'une ontologie. Il s'agit des outils Virtuoso [5] et D2RQ [6]. Dans notre approche nous avons choisi d'utiliser D2RQ.

D2RQ est une plateforme de publication de BDR sur le Web utilisant les standards du Web Sémantique et permettant de traiter une base de données relationnelle comme un graphe virtuel RDF. Dans ce graphe un élément du schéma est représenté par un nœud et une relation par un arc orienté. Il est possible de créer ce graphe virtuel RDF en exportant uniquement le schéma de la base de données relationnelle et donc aucune instance. Nous parlerons alors de vue RDF. La plateforme D2RQ est composée de 3 éléments principaux. i) Le langage de mapping D2RQ qui est un langage déclaratif utilisé pour créer la vue RDF de la BDR et permettant de décrire les relations entre des ontologies et un schéma de BDR. ii) Le moteur D2RQ permettant de créer automatiquement une vue RDF et de transformer une requête SPARQL interrogeant la vue RDF en une requête SQL interrogeant directement la BDR. iii) le Serveur D2R qui est un serveur HTTP permettant d'interroger les bases de données relationnelles via le Web.

## 3 BioSemantic

BioSemantic [3] est une plateforme d'intégration de BDR de type médiation basée sur les standards du Web Sémantique découpant la création d'adaptateurs en 2 étapes distinctes. Une première étape consiste à créer automatiquement une vue RDF du schéma de la BDR à intégrer, puis à annoter manuellement cette vue à l'aide de termes ontologiques. L'étape d'annotation est la seule nécessitant un utilisateur expert ayant une connaissance du schéma de la BDR à intégrer. La 2ème étape est l'étape de création d'adaptateurs à proprement parler. Elle utilise toutes les vues RDF précédemment créées et annotées pour créer automatiquement des adaptateurs. Dans cette seconde étape aucune connaissance des schémas de BDR n'est nécessaire. La seule nécessité est la connaissance des termes ontologiques utilisés dans le schéma global. La création de ces adaptateurs est basée à la fois sur un enrichissement sémantique de la vue RDF créée par D2RQ et sur la notion de parcours de graphe.

### 3.1 Intérêts de D2RQ dans BioSemantic

L'utilisation de D2RQ dans BioSemantic présente plusieurs avantages comme la présence d'un langage déclaratif permettant de définir des mappings complexes entre le schéma de la BDR et des termes ontologiques. Un autre avantage est la présence du moteur D2RQ permettant de transformer automatiquement une requête SPARQL interrogeant une vue RDF en une requête SQL interrogeant la BDR sans exporter ses données. De plus, l'utilisation de RDF et son formalisme en triplet va nous permettre de parcourir la vue RDF comme un graphe et ainsi virtuellement parcourir le schéma de la BDR pour trouver une requête pertinente à intégrer dans notre adaptateur. Nous allons donc utiliser D2RQ tout en détournant son utilisation pour automatiser la création de requêtes SPARQL.

### 3.2 Limites de l'utilisation de D2RQ dans BioSemantic

Nous souhaitons utiliser le langage D2RQ pour parcourir notre vue RDF et ainsi indirectement parcourir notre schéma de BDR. D2RQ n'ayant pas été implémenté pour cette utilisation, le langage D2RQ n'est pas suffisamment expressif pour définir toutes les relations que nous souhaiterions. Il permet par exemple de définir des clés primaires et des clés étrangères mais ne permet pas de définir une table d'association ou d'autres spécificités à prendre en compte pour notre parcours de la vue RDF dans le but de créer automatiquement une requête pertinente. Une requête créée automatiquement sera considérée comme pertinente si les données qu'elle renvoie sont identiques aux données renvoyées par une requête créée

manuellement par un expert du schéma de la BD. Dans cet article nous présenterons uniquement l'enrichissement sémantique de la vue RDF D2RQ dans le but de créer automatiquement une requête. L'algorithme de parcours de la vue RDF ne sera pas détaillé.

## 4 Enrichissement sémantique d'une vue RDF D2RQ

Lors de la création de requête, il faut tenir compte du contexte spécifique de la recherche de plus court chemin dans un schéma de base de données relationnelle. Une des spécificités de notre schéma relationnel par rapport à un simple graphe, est la présence de relations bien définies entre les tables. Notre recherche de plus court chemin doit donc tenir compte des spécificités de ces relations.

### 4.1 La combinaison de chemins

#### a) Relations concernées

**La relation d'agrégation:** par défaut, une association exprime une relation à couplage faible. Les entités associées restent relativement indépendantes l'une de l'autre [7]. L'agrégation est une forme particulière d'association qui exprime un couplage plus fort entre entités. Elle permet d'exprimer des relations de type maître/esclaves et représentent des connexions bidirectionnelles dissymétriques.

**La relation de composition:** il s'agit d'une forme d'agrégation avec couplage plus important entre les entités. Cette composition indique que la destruction de l'agrégat entraîne automatiquement la destruction des composants agrégés.

**La relation d'héritage:** la généralisation et la spécialisation sont des points de vue portés sur les hiérarchies d'entités. Une entité A est une spécialisation d'une entité B si chaque instance de A est une instance de B et si chaque instance de B est associée à au plus une instance de A.

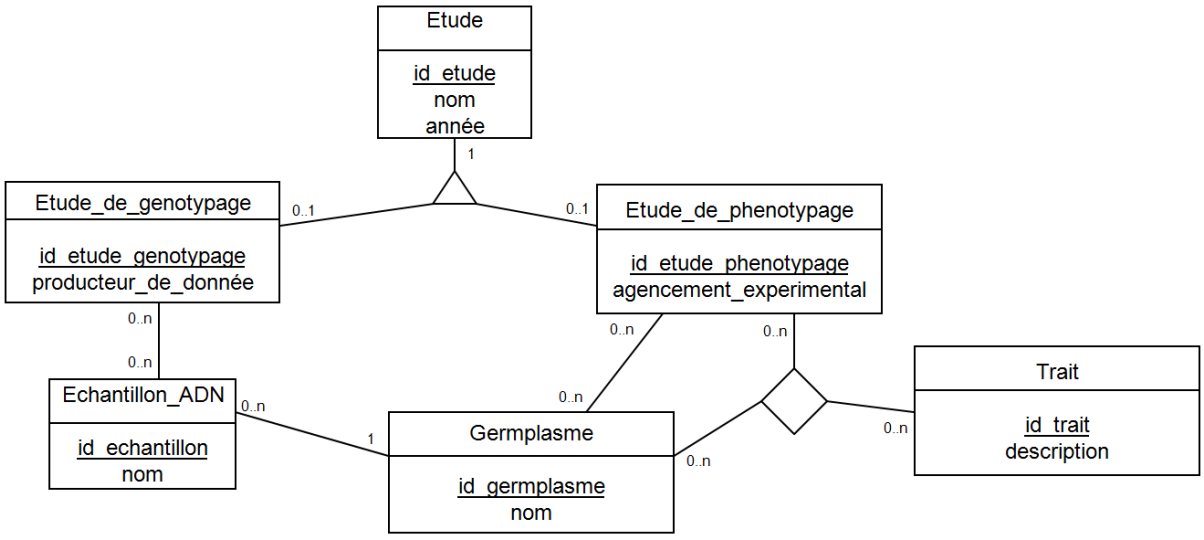



**Figure 1.** Schéma entité-association montrant une relation d'héritage

Nous nous sommes intéressés aux relations d'agrégation et d'héritage car elles présentent une particularité commune susceptible de nous intéresser. Toutes les entités spécialisées, issues d'une même relation d'héritage, sont indépendantes. Cela signifie que l'intersection des instances qu'elle contient est vide. Cette propriété est également présente dans les entités agrégées issues d'un même agrégat et pourrait être utilisées pour combiner différents chemins pour créer notre requête sans que cette dernière ne renvoie des données redondantes. Nous allons illustrer l'intérêt de la combinaison de chemins à l'aide de l'exemple présent dans le schéma d'entité-association de la Figure 1 ci dessous. Dans ce schéma on peut voir une relation d'héritage. Les entités *etude_de_genotypage* et *etude_de_phenotypage* sont des spécialisations d'une *etude*. Une étude est donc soit une étude de génotypage, soit une étude de phénotypage. Cela peut poser des

problèmes si on souhaite créer automatiquement une requête qui, pour un identifiant d'étude donné, renvoie tous les noms de germplasmes associés. En utilisant un algorithme de recherche de plus court chemin, seul le chemin ayant le moins de nœuds intermédiaires permettant de relier *etude* à *germplasme* serait détecté. Dans notre exemple le chemin détecté serait le suivant:

etude – etude_de_phenotypage - germplasme

Ce chemin serait utilisé pour créer une requête composée de jointure entre ces tables. Cela signifie que si un identifiant donné correspond à un identifiant d'étude de génotypage, la requête créée automatiquement ne renverra aucune donnée. Dans le cas de relations d'héritage dans un schéma entité-association, il faut donc utiliser une combinaison de chemins pour la création de requêtes. Cependant, pour savoir comment prendre en compte ce type de relation, il faut s'intéresser aux règles de conversion de ce type de relation du modèle entité-association au modèle relationnel.

**b) Passage au modèle relationnel**

La conversion d'une relation d'héritage peut s'effectuer de trois façons différentes lors du passage du modèle entité-association vers le modèle relationnel [8] aplati vers le haut, aplati vers le bas ou non aplati. Les relations d'héritage aplaties vers le haut et vers le bas ne posent pas de problème de combinaison de chemin lors de notre recherche de plus court chemin. En suivant une approche non aplatie, chaque entité est convertie en une relation. Une clé étrangère supplémentaire, correspondant à la clé primaire de la relation généraliste, est présente dans chaque relation spécialisée. Le modèle résultant est le suivant:

*etude(id_etude, nom, annee)*
*echantillon_ADN(id_echantillon, nom, #id_germplasme)*
*germplasme(id_germplasme, nom)*
*etude_de_genotypage (id_etude_genotypage, producteur_de_donnee, #id_etude)*
*etude_de_phenotypage(id_etude_phenotypage, agencement_experimental, #id_etude)*
*echantillon_genotypage (#id_echantillon, # id_etude_genotypage)*
*germplasme_phenotypage (# id_germplasme, # id_etude_phenotypage)*

Ici les attributs soulignés correspondent à des clés primaires et les attributs précédés d'un dièse correspondent à des clés étrangères. Lors d'une conversion vers une relation d'héritage non aplati, les relations spécialisées et généralistes sont présentes physiquement. Cela signifie qu'une relation est créée pour chaque entité. Dans ces conditions, la création d'une requête prenant en entrée la relation généraliste peut nécessiter de combiner plusieurs chemins. Pour la conversion non aplatie de la relation d'héritage, présentée dans notre exemple, la création d'une requête renvoyant tous les germplasmes pour une étude donnée nécessitera la combinaison des plus courts chemins passant par les relations *etude_de_genotypage* et *etude_de_phenotypage*.

**c) Détection des relations d'héritage non aplaties**

La détection automatique de relations d'héritage non aplati pour la transformation d'un schéma relationnel vers une ontologie a été décrite dans [9]. Elle est également utilisée pour typer les relations d'héritage de l'outil DB2OWL [10]. Cette détection automatique est basée sur les techniques d'ingénierie inverse dans les bases de données, tentant notamment de convertir un modèle relationnel en modèle entité association [11]. Cette détection utilise la particularité des contraintes d'intégrité entre les tables généralistes et les tables spécialisées. En effet, pour qu'une table soit une spécialisation d'une autre table, elle doit contenir pour seule clé étrangère la clé primaire de la table généraliste. La détection automatique de ce genre de relation d'héritage est donc rendue possible en utilisant la règle suivante:

***Subclass(r,s)*** *<-* ***Rel(r)^Rel(s)^PK(x,r)^FK(x,r,_,s)***
*avec*
*Rel(r) r est une relation*
*PK(x,r) x est la clé primaire de r*
*FK(x,r,y,s) x est la clé primaire de la relation r et référence y dans la relation s*

L'utilisation de cette règle rend automatique la détection de toutes les relations d'héritage non aplaties. Elle détecte également les relations d'agrégation ou de composition nous permettant donc de détecter tous les types de chemins que nous souhaitions combiner lors de la création de nos requêtes.

#### d) Enrichissement de la vue RDF

Nous avons décidé d'ajouter cette information directement dans la vue RDF, à l'aide de la propriété *rdfs:SubClassOf* issue du vocabulaire RDF. Dans notre exemple, la détection de relation d'héritage implique la création de deux nouveaux triplets dans la vue RDF.

| | | |
|---|---|---|
| *etude_de_genotypage* | *rdfs:subClassOf* | *etude* |
| *etude_de_phenotypage* | *rdfs:subClassOf* | *etude* |

### 4.2 Prise en compte de la pondération dans la sélection de chemin

La simple prise en compte du nombre de nœuds à parcourir pour déterminer la longueur du plus court chemin ne prend pas en compte la diversité d'associations présentes dans un schéma relationnel. Dans notre vue RDF nous n'avons pas accès aux cardinalités des associations. Nous ne pouvons donc pas utiliser ce critère pour favoriser le passage par une association plutôt que par une autre. Nous avons par contre accès aux clés primaires et clés étrangères. Nous allons donc utiliser ces paramètres pour tenter d'optimiser l'algorithme de détection de plus court chemin. Nous allons dans un premier temps nous intéresser aux tables d'associations puis aux arités des relations.

#### a) Prise en compte des tables d'association

Lors du passage du schéma entité association au schéma relationnel, une entité devient une relation contenant des attributs. Les associations binaires, de type 1:1 ou 1:n disparaissent au profit d'une clé étrangère dans la relation coté 0,1 ou 1,1. Cette clé étrangère référence la clé primaire de l'autre relation. Une association binaire de type n:m, est convertit en une relation qui, dans le modèle physique, correspondra à une table d'association. La clé primaire de cette table est alors composée de 2 clés étrangères référençant les 2 clés primaires des 2 tables associées. Dans notre contexte de recherche de plus court chemin reliant 2 nœuds d'un graphe, il serait peut être intéressant de ne pas traiter ce type de table de la même façon qu'une table représentant une entité.

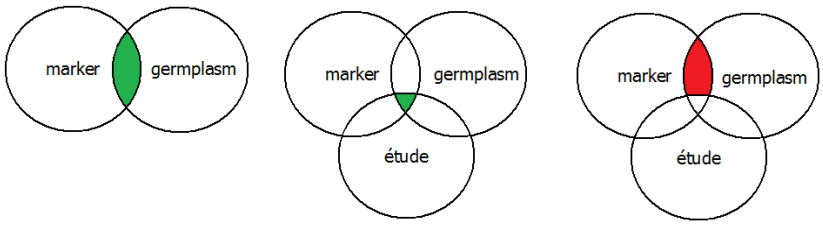



**Figure 2.** Représentations des données contenues dans 2 tables d'association d'arité différente

L'association binaire, est schématisée dans la partie gauche de la Figure 2. La partie verte représente les instances contenues dans la table d'association.. La partie centrale de la Figure 2 schématise quand à elle une association ternaire. La zone verte représente les instances contenues dans la table d'association résultante. La prise en compte d'une 3ème association dans la table d'association conduira à une perte d'information présentée en rouge dans la partie droite de la Figure 2. Le passage par des tables d'association d'arité supérieure à 2 peut induire une perte d'information. Nous avons donc intérêt à favoriser le passage par les tables d'association binaires mais également de pénaliser le passage par les tables d'association d'arité supérieure à 2.

#### b) Détection des tables d'association et de leur arité

La Figure 3 présente l'algorithme de détection des tables d'association avec attributs.

*pk= clé primaire de R*
*fk=clés étrangères de R*
$if\ ((\forall u \in R)(u \in fk => u \in pk))\{$
$if((\forall u \in R)(u \in pk => u \in fk)\{$
*R est une table d'association*
*}*
*}*

**Figure 3.** Algorithme de détection des tables d'association avec attributs

Une table d'association sans attributs est un cas particulier de table d'association. Il s'agit d'une table associant 2 autres tables et dont les seules colonnes sont celles correspondant aux clés primaires des 2 tables associées. Dans la Figure 4**,** c'est le cas de la table d'association *germplasme_phenotypage* créée suite au passage du schéma entité association de la Figure 1 vers un modèle relationnel sans aplatir les relations d'héritage. Ce type de table d'association est typé comme une propriété dans la vue RDF D2RQ. La détection de ces tables d'association sans attributs se fait automatiquement en détectant les propriétés de la vue RDF possédant plusieurs clés étrangères.

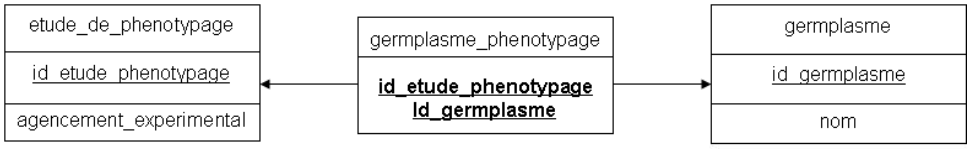



**Figure 4.** Exemple de table d'association sans attributs

Dans la vue RDF, l'information sur l'arité d'une table d'association est transcrite indirectement. Cette arité peut être trouvée automatiquement en détectant le nombre de clés étrangères présentent dans une table d'association.

#### c) Enrichissement de la vue RDF D2RQ

Les tables d'associations possédant ou non des attributs sont annotées avec la propriété *dr:associatedTo*. Un triplet contenant ce prédicat sera ajouté pour chaque table associée. Le sujet de ce triplet correspondra à la table d'association et l'objet à la table associée. L'arité d'une table d'association est annotée avec la propriété *dr:arity*. Ce typage est réalisé automatiquement, sous la forme de triplets, lors de la création de la vue RDF. Dans les exemples de la Figure 4, les triplets suivants sont ajoutés à la table d'association binaire nommée *markergermplasm* associant la table *marker* et la table *germplasm*:

*map:germplasme_phenotypage* *dr:associatedTo* *map:germplasme*
*map:germplasme_phenotypage* *dr:associatedTo* *map:etude_de_phenotypage*
*map:germplasme_phenotypage* *dr:arity* *"2"^^rdf:int*

## 5 Résultats

### 5.1 Utilisation de l'enrichissement dans la recherche de plus court chemin

Ce schéma a conduit à la création d'une vue RDF dont la représentation sous forme de graphe est présentée dans la Figure 5. Dans ce graphe, seul les nœuds représentant des tables sont présents, ces nœuds sont de couleur orange. Les nœuds rouges représentent les annotations sémantiques, ajoutées manuellement, réalisées sur une colonne de la table associée. Les arcs noirs représentent les propriétés présentes d'origine dans la vue RDF, et les arcs rouges représentent les arcs rajoutés automatiquement par notre approche. Les nœuds bleus représentent la valeur associée à l'arité d'une table d'association, qui est détectée automatiquement. Pour créer une requête SPARQL, nous utilisons les annotations sémantiques ajoutées manuellement dans la vue RDF d'un schéma de base de données relationnelle. Dans l'exemple de la Figure 5, nous allons sélectionner les annotations sémantiques *gcpdm:etude* et *gcpdm:germplasm* pour créer automatiquement une requête renvoyant tous les germplasms d'une étude donnée.

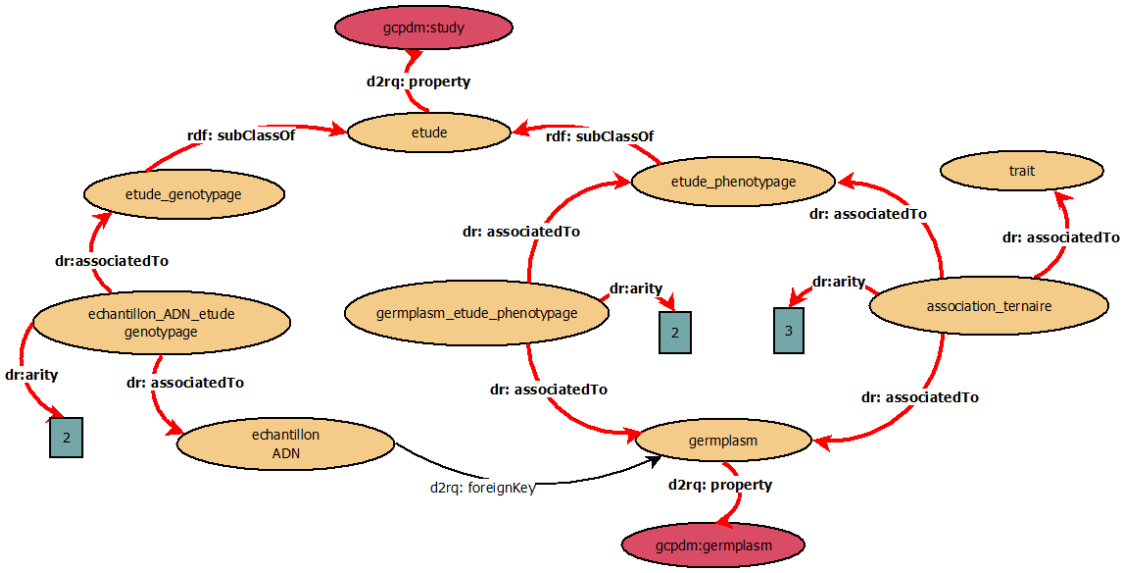



**Figure 5.** Graphe représentant la vue RDF du schéma de la base de données utilisée comme exemple de la création de requête SPARQL

Lors de la recherche de plus court chemin, l'algorithme va détecter une relation de spécialisation entre la table *etude* et les tables *etude_genotypage* et *etude_phenotypage* grâce à l'enrichissement sémantique avec les balises rdf:subclassOf. Cette information va être prise en compte et le plus court chemin renvoyé sera donc l'agrégation des plus courts chemins passant par ces 2 tables spécialisées (flèches rouges de l'étape A de la Figure 6). Lors du passage par la table *etude_phenotypage*, l'algorithme a la possibilité de trouver 2 chemins passant par le même nombre de noeuds. Le premier chemin passe par la table d'association binaire *germplasm_etude_phenotypage* et le deuxième chemin par la table d'association ternaire appelée ici *asociation_ternaire*. L'enrichissement sémantique avec les balises dr:associatedTo permet à l'algorithme de détecter l'arité de ces tables d'association et de choisir de passer par celle ayant l'arité la plus petite. Dans l'exemple de l'étape B de la Figure 6, l'algorithme passera par la flèche rouge de gauche et ne parcourra pas la portion de graphe passant par la flèche rouge droite. Le plus court chemin final renvoyé par notre algorithme et permettant de créer automatiquement une requête pertinente est représenté dans l'étape C de la Figure 6.

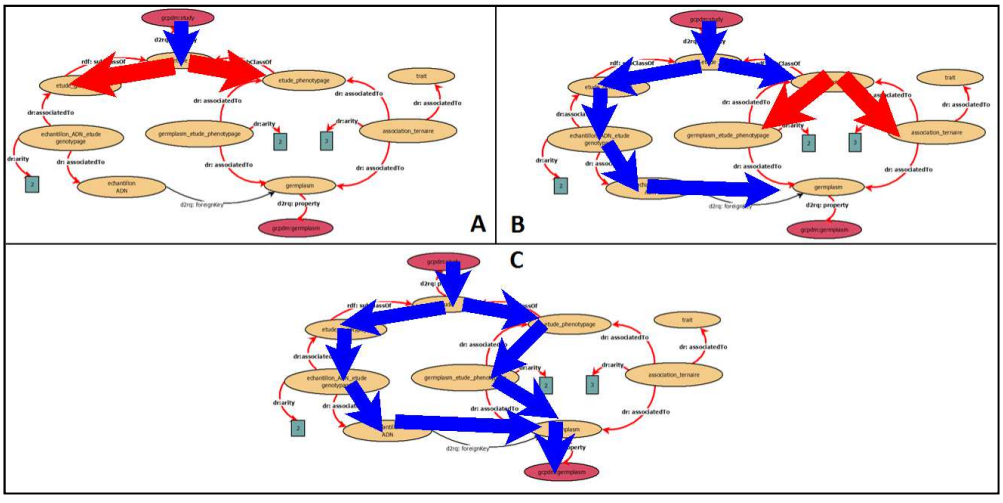



**Figure 6.** Utilisation de l'enrichissement sémantique dans le parcours de graphe

## 5.2 Pertinence des requêtes

Nous avons comparé, dans la Table 1, les données renvoyées par 3 approches de création de requêtes différentes: i) un algorithme de Dijkstra récupérant le plus court chemin, ii) BioSemantic et iii) une requête

SQL créée manuellement par un utilisateur expert. La pertinence de nos requêtes a été testé sur un grand nombre de requêtes. La Table 1 représente les données renvoyées par les 3 approches selon la présence de relation d'héritage et/ou de la présence de plusieurs tables d'association d'arité différentes. Pour cet exemple, nous avons utilisé la BDR d'espèces cultivées tropicales TropGene [12].Les requêtes créées automatiquement par BioSemantic renvoient le même nombre de résultats que les requêtes créées manuellement par l'utilisateur expert. Dans le cas de l'algorithme de Dijkstra la quantité de données renvoyées n'est pas toujours la même que celle des requêtes créées manuellement. Cela est dû au fait qu'il ne prenne pas en compte les relations d'héritage et les cas ou des tables d'associations d'arité différentes relies 2 mêmes tables.

| | **héritage** | **Plusieurs tables d'association d'arité différentes** | **Dijkstra** | **BioSemantic** | **Requête SQL manuelle** |
|---|---|---|---|---|---|
| Requête 1 | Oui | non | 1595 | 7212 | 7212 |
| Requête 2 | Non | oui | 0 | 12302 | 12302 |
| Requête 3 | Non | oui | 197 | 197 | 197 |
| Requête 4 | Non | non | 2055 | 2055 | 2055 |

**Table 1.** Comparaison des données renvoyées selon 3 approches: l'algorithme de Dijkstra, BioSemantic, et une requête SQL créée manuellement.

## 6 Conclusion

Nous avons développé un outil nommé BioSemantic facilitant l'interrogation de bases de données relationnelles distribuées en automatisant la création de requêtes SPARQL. Nous avons présenté la manière dont nous avons détourné l'utilisation classique de D2RQ, un outil de mise en correspondance entre bases de données relationnelles et ontologies, afin d'automatiser au maximum la création de requêtes distribuées. Pour cela nous avons enrichi sémantiquement la vue RDF créée automatiquement par D2RQ. Les métadonnées ainsi ajoutées permettent de détecter les relations d'héritages et de composition, les arités entre tables ainsi que de typer les tables d'associations. Les vues RDF enrichis peuvent ainsi être utilisées dans BioSemantic pour automatiser la création de requêtes pertinentes sur des bases de données relationnelles distribuées.

## Références

[1] C. Goble et R. Stevens, « State of the nation in data integration for bioinformatics », *J Biomed Inform*, vol. 41, n° 5, p. 687‑693, oct. 2008.

[2] E. Antezana, M. Kuiper, et V. Mironov, « Biological knowledge management: the emerging role of the Semantic Web technologies », *Briefings in Bioinformatics*, vol. 10, n° 4, p. 392 ‑407, juill. 2009.

[3] J. Wollbrett, P. Larmande, F. de Lamotte, et M. Ruiz, « Clever generation of rich SPARQL queries from annotated relational schema: application to Semantic Web Service creation for biological databases », *BMC Bioinformatics*, in press.

[4] D.-E. Spanos, P. Stavrou, et N. Mitrou, « Bringing Relational Databases into the Semantic Web: A Survey », *Semantic Web*, vol. 3, n° 2, p. 169‑209, 2012.

[5] O. Erling et I. Mikhailov, « Mapping Relational Data to RDF in Virtuoso », *OpenLink Software*, 2006. [Online]. Available: http://virtuoso.openlinksw.com/wiki/main/Main/VOSSQLRDF. [Accessed: 07-sept-2011].

[6] C. Bizer, « D2RQ - treating non-RDF databases as virtual RDF graphs », *IN PROCEEDINGS OF THE 3RD INTERNATIONAL SEMANTIC WEB CONFERENCE (ISWC2004*, 2004.

[7] Pierre-Alain Muller et Nathalie Gaertner, *Modélisation objet avec UML*, Deuxième édition. Eyrolles, 2000.

[8] R. Elmasri, *Fundamentals of database systems*, 5. ed. Boston: Pearson, 2006.

[9] S. Tirmizi, J. Sequeda, et al., « Translating SQL Applications to the Semantic Web », in *Database and Expert Systems Applications*, 2008, p. 450‑464.

[10] N. Cullot, R. Ghawi, et K. Yétongnon, « DB2OWL : A Tool for Automatic Database-to-Ontology Mapping », presented at the SEBD, 2007, p. 491‑494.

[11] K. H. Davis et A. K. Arora, « Converting A Relational Database Model into an Entity-Relationship Model », in *Proceedings of the Sixth International Conference on Entity-Relationship Approach*, 1987.

[12] M. Ruiz, M. Rouard, L. M. Raboin, M. Lartaud, P. Lagoda, et B. Courtois, « TropGENE-DB, a multi-tropical crop information system », *Nucleic Acids Res.*, vol. 32, n° Database issue, p. D364‑367, janv. 2004.

# A Small Step into Galaxy, a Faster Pace for Metabolomics

## Galaxy and the metabolomics analysis Universe

Pierre PERICARD[1], Gildas LE CORGUILLE[1], Urszula CZERWINSKA[1], Marion LANDI[2], Franck GIACOMONI[2], Christophe DUPERIER[2], Jean-François MARTIN[2], Sophie GOULITQUER[1], Estelle PUJOS-GUILLOT[2], and Christophe CARON[1]

[1] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France
{pierre.pericard, gildas.lecorguille, ursula.czerwinska, sophie.goulitquer, christophe.caron}@sb-roscoff.fr

[2] PFEM, UMR1019 INRA, Centre Clermont-Ferrand-Theix, 63122, Saint Genes Champanelle, France
landi.marion@gmail.com
{franck.giacomoni, christophe.duperier, jean-francois.martin, estelle.pujos}@clermont.inra.fr

***Abstract*** *Facing the emergence of new technologies in the field of metabolomics, treatment solutions adopted so far (XCMS, R scripts, etc.) clearly show their limits. Bottlenecks affect unified access to core applications as well as computing infrastructure and storage. In the context of collaboration between metabolomics and bioinformatics platforms, we have developed a full pipeline using Galaxy framework for data analysis. This modular and extensible workflow includes existing components (XCMS functions, etc.) but also a whole suite of complementary statistical tools. This implementation is accessible through a web interface, which guarantees the parameters completeness. The advanced features of Galaxy have made possible the integration of components from different sources and of different types. Finally, an extensible environment is offered to the metabolomics community, and enables preconfigured workflows sharing for new users, but also experts in the field.*

**Keywords** Metabolomics, Galaxy, XCMS, workflow

## Un Petit Pas dans Galaxy, et la Métabolomique s'Accélère

### Galaxy et l'Univers des analyses métabolomiques

**Résumé** *Face à l'arrivée de nouvelles technologies dans le domaine de la métabolomique, les solutions de traitements adoptées jusqu'à maintenant (XCMS, scripts R, etc.) montrent clairement des limites. Les verrous concernent aussi bien l'accessibilité unifiée aux applications métiers que les problèmes d'infrastructure de calcul ou de stockage. Dans le cadre d'une collaboration entre les plateformes INRA/PFEM et CNRS/ABiMS-METABOMER, nous avons développé sous Galaxy un pipeline complet d'analyse. Ce workflow modulaire et extensible, inclut des composants existant (fonctions XCMS, etc.) mais aussi toute une suite d'outils statistiques complémentaires. Cette implémentation, accessible au travers d'une interface web, garantie l'exhaustivité des paramètres. Les fonctionnalités avancées de Galaxy ont permis l'intégration de composants provenant de différentes sources et de nature différente. Au final, un premier environnement est proposé à la communauté métabolomique, et permet le partage de workflows préconfigurés à destination d'utilisateurs novices, mais aussi d'experts du domaine.*

**Mots-clés** Métabolomique, Galaxy, XCMS, workflow

## 1 Introduction

The development of "omics" sciences give biologists access to the large variety of the life components and allows the study of metabolism reflecting a possible integration of all post genome phenomena. Recent advances in analytical tools such as high performance liquid chromatography coupled with mass spectrometry (LC-MS) and efforts in chemometrics and in biological computing allows imagining new

analysis strategies to study chemical processes involving metabolites. Metabolomics can be described as a global analysis of small molecules of a biological sample, which are produced or modified as a result of stimuli [1]. The production and utilization of metabolites seems to be more directly connected to the phenotype exhibited by an organism than the presence of mRNAs or proteins.

Metabolomic experiments include different steps based on several technical knowledge i) sample preparation, ii) metabolic profiling, iii) extraction and alignment of data, iv) statistical treatment of data set, v) identification of key or discriminatory metabolites vi) their quantification. This approach has the particularity of generating a large amount of massive datasets. Several thousand of features are produced in several hundred of samples. The extraction of ions from acquisition files is a key step in metabolomic studies, as extraction and alignment software packages provide dataset used for further statistical analysis and identification of metabolites. Alignment of multiple samples may produce noise in the list of extracts ions with problem of overfitting model. So, the choice of software settings is crucial and must be consider as one the most important step in the metabolomic data workflow [2].

Over the past decade, a lot of algorithms and tools were proposed by the scientist community (XCMS, MzMine [4], etc…) and by commercial companies. These applications are standalone software programs or command line scripts, in freeware or commercial versions, with really specific functionalities and often limited performance. Since 2009, some web analysis applications like XCMS Online [5] and MetaboAnalyst [6] had been implemented with a particular effort on user interface compared to other described solutions. As an alternative solution, we propose a tool box, easy to grow, functionality by functionality, module by module, from multi-lab sources, accessing by simple web interface and adapted to be used by biologists, MS-experts or statisticians or "bioinformaticians". Furthermore, this project fits into a method adaptability context and up-scaling of labs in silico analysis ability.

The web open-source analysis platform Galaxy (http://galaxyproject.org/) was chosen to integrate a suite of current applications in metabolomic community. Galaxy is one of the most useful systems compared to other workflow engines and seems to be adapted to deal with metabolomic data and analysis: no known data size limitations [10], possibilities to automate pipelines, and to ensure reproducibility. Because of its web interface, the system allows true cross platform availability and runs analysis chain by scientists without programming experience.

A metabolomic version of Galaxy including the XCMS suite and statistical modules (Hierarchical clustering, PCA, Normalization) was implemented with the objective to establish a proof of concept based on the new platform adaptability, the interfacing of the original XCMS modularity, and validating ergonomics in comparison to other versions of this framework (R package and XCMS Online versions). This project aims to be the first step of a more ambitious project consisting in the building of complete analysis workflows with extraction, identification and biological interpretation modules.

## 2 Materials and Methods

### 2.1 XCMS / CAMERA

XCMS [3] is a free R package which allows extraction and quantification of ions obtained by liquid chromatography coupled with mass spectrometry. XCMS operates on raw acquisition data files converted into NetCDF, mzXML or mzData format. These formats can be produced by all kind of mass spectrometer softwares. XCMS aims to provide a peak list in four steps corresponding to four R functions. Firstly, ions are extracted from each sample independently. Using a Gaussian model, peaks are filtered and integrated. In a second step, called alignment or grouping, individual peaks are matched across all the samples. The third step is an optional correction of retention time drift using a non-linear LOWES regression. Then, if one observed a significant correction of retention time it is necessary to run the second step again. The last step also optional is a gap filling in order to replace missing data by baseline noise. The modularity of these four steps is an attractive feature of XCMS. When the peak list is defined, the diffreport function performs univariate statistics between the different studied conditions and proposes ions annotation based on METLIN database. Finally, the CAMERA package can be used to identify adducts or fragment ions generated in the mass spectrometer ionization source which are redundant information.

## 2.2 Galaxy

Our metabolomics tools were integrated to the open, web-based platform: Galaxy [7,8,9]. Initially developed for the computational biomedical research community, its application spectrum has grown increasingly wider. Therefore, more and more bioinformatics platforms, in France and around the globe, have adopted it. Galaxy provides an interface for tools and workflows, and is designed to be: accessible – users without programming experience can easily specify parameters and run tools and workflows; reproducible – Galaxy captures information so that any user can repeat and understand a complete computational analysis; transparent – users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis. Galaxy very active community also insures up-to-date software releases and efficient support for both end-users and tools developers.

## 2.3 Infrastructure

The Galaxy platform dedicated to metabolomics data analysis was integrated to the standard components of ABiMS (Analysis and Bioinformatics for Marine Science) platform computing infrastructure. Galaxy was deployed in a virtual environment based on VMWARE. Optimization efforts, such as connections pool or web services decoupling, allow a good level of scalability. Computing resources connection uses the standard DRMAA API, and is completed with a dedicated connector (tool runner) in order to make available adequate resources both in terms of high computing performance and memory amount ( up to 1TB RAM). Finally, a shared and secure storage space completes this layer, essential to smoothly working treatments.

## 2.4 Inputs

The XCMS package can read full-scan LC/MS data from NetCDF, mzXML, and mzData files. A single experiment can generate from tens to several hundreds files, which have to be organized in a specific arborescence so that the distinct conditions can be identified by XCMS. To easily handle such inputs, we added to our Galaxy instance a new proprietary datatype based on a zip file but with a specific extension (.ms.zip). To prevent Galaxy to unzip these files we also patched the “Get Data” tool and added a sniffer specific to our new datatype. These modifications allow .ms.zip files to be uploaded and automatically assigned the correct datatype with no manual intervention from the user. To allow communication between our pipeline tools, we also added a second datatype based on RData files, which saves the information from one tool and can be used as the input of the next tool.

## 2.5 Galaxy Tool XML Definition Files

The first step to integrate a tool in Galaxy consists in creating its XML definition file (cf. Figure 1), which let Galaxy know the execution details of the new tool. Therefore, we wrote nine definition files (one for each step of our pipeline) according to the good development practices guidelines from the IFB (Institut Français de Bioinformatique) Galaxy working group. In order to provide to XCMS users an ergonomic interface, the large majority of the tool parameters were described, typed, and discriminated between main and advanced parameters. We also leveraged the Galaxy conditional system to dynamically display parameters depending on other parameters. Every tool and parameter was documented based on the official documentation and our implementation specificities.

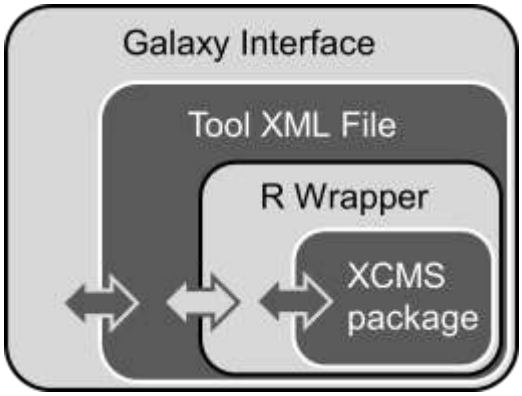



**Figure 1.** Galaxy multiple layers organization

### 2.6 R Wrapper and XCMS Functions

Galaxy XML wrappers cannot directly run R functions and/or multiple commands in a single run. Therefore, to run R functions we need to develop wrappers in R (cf. Figure 1). We chose to write only one script/wrapper by package (in this case XCMS/CAMERA) and keep it as inclusive and transparent as possible. As a consequence, adding a parameter or function (eg. following a XCMS update) only require some changes in the XML definition file. This script allows libraries loading, argument passing between the XML definition files and the R functions, and data preprocessing.

### 2.7 Post Extraction Normalization and Statistical Analyses

After peaklist definition, the pipeline continues with normalization and a series of univariate and multivariate statistical analyses, available using R functions. During LC-MS analyses, especially with large number of samples, the analysis is often interrupted for mass spectrometer maintenance or calibration. Moreover, it is also observed a decrease of intensity due to ionization source clogging during a batch analysis. This leads to analytical intensity drifts. It is possible to correct this evolution of intensity by normalization based on linear regression [11]. In order to select metabolites with significant intensity difference among the different experimental conditions, an analysis of variance can be carried out with Benjamini Hochberg p-values correction (BH). Defining a p-value threshold, significant ions are selected for further unsupervised multivariate analyses. Hierarchical Ascendant Clustering using hclust R function (using different options of distance and aggregation methods) can be used to aggregate ions depending on their abundance in samples. A result file is produced in order to use Treeview [12] software for interactive clustering of sample and metabolites and for heatmap construction. Principal Component Analysis (PCA) is also available using FactoMineR (http://factominer.free.fr/) R package.

## 3 Results and Discussion

A total of nine XCMS R functions and statistical analysis scripts were implemented in Galaxy as a metabolomics analysis pipeline (cf. Figure 2). Each tool can be run independently or as part of a complete workflow. We also successfully implemented the parallelization of the supported XCMS functions. The “Rmpi” R library was installed on the ABiMS cluster and all the related Galaxy wrappers were configured to use the XCMS multi-core options. This configuration allows to greatly speed up some of XCMS most time-consuming functions (xcmsSet, by ex). MPI runs also allow a smoother cluster usage while saving time during CPUs reservation (vs. a usual multi-thread run).

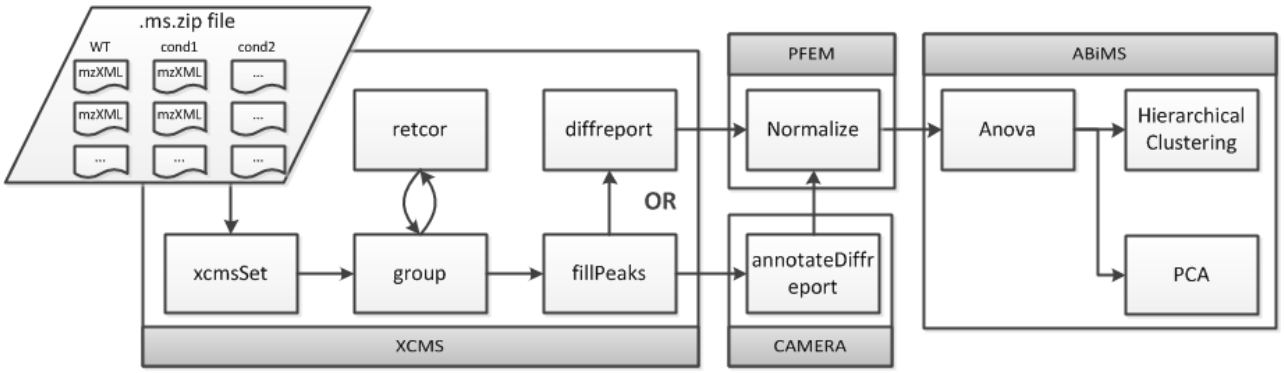



**Figure 2.** Organization of our metabolomic analysis workflow components.

In order to assess the different tools and verify that they all leads to the same peaklist, a real set of samples using Netcdf files format was used to compare the extraction results using XCMS command line (XCL), XCMS Online (XO) and XCMS implemented in Galaxy (XG). As expected almost the same list of ions was detected: 911 ions for XCL, 912 for XG and 962 for XO with an average coefficient of variation of peak integration of 0.0014%. These results confirm that we obtained very close results with these different techniques.

XCMS for R command line is the reference software. It has a great potential when associated with existing R packages especially for statistical analysis. Though programming makes possible defining the workflow and integrating in-house scripts for normalization and annotation of ions with public databases, it highly impacts ergonomics. Mass spectrometer analysts are generally not R users and so there is a need of interface to easily use the XCMS package. XCMS Online and Galaxy are both very attractive alternatives to XCMS for R as they are both based on a web interface. XCMS Online is dedicated to non XCMS users by providing pre-configured workflows with parameters for each kind of mass spectrometers. However, some experts can be limited by the lack of parameter completeness. Instead, our implementation of XCMS in Galaxy was designed to display as much parameters as the R functions, although non-XCMS users have access to pre-configured workflows shared by the community.

| | **R** | **XCMS Online** | **XCMS on Galaxy** |
|---|---|---|---|
| **Ergonomics** | -- | ++ | + |
| **Parameter Completeness** | ++ | - | + |
| **Target** | R and XCMS advanced users | Non-XCMS users<br>XCMS advanced users - | Non-XCMS users<br>XCMS advanced users +<br>Additional tools developers |
| **Data & workflow sharing** | - | - | ++ |
| **Modularity** | + | - | ++ |

**Table 1.** XCMS implementation comparison

Modularity features of Galaxy, which allows integrating multi languages applications (Python, Perl, R, Matlab, Bash, C, etc.) from multi repositories (toolshed or others), was used in this project to build components and workflows in two geographical places and in two teams of our community: ABIMS and PFEM platforms. Galaxy plays a central role in merging software engineering methods and sharing private tools (cf. Table 1).

## 4 Conclusion and Perspectives

To face the predicted fast and large request production of metabolic fingerprints, scientist communities get organized in data representation standards development and in primary database models implementation. In this context, the ability of biological computing actors, proposing in-silico strategies and analysis tools, is the key factor of the future of metabolomics like a high-speed science, as well as the necessary advances in mass-spectrometry technologies.

We successfully implemented and made available a full bioinformatics pipeline for metabolomics analysis. This pipeline can be used through the ABiMS Galaxy web interface, which provides an easy way to test, parameter, and keep a history of previous analysis. Workflows can also be designed by combining several tools. Experts can pre-configure these workflows using advanced parameters and then make them available to all users through the Galaxy sharing interface. To illustrate this part, the ABiMS Galaxy instance already proposes two expert and pre-configured workflows for metabolomics analysis with either medium or high resolution mass-spectrometers.

With this national collaboration between two metabolomics platforms PFEM and METABOMER, and the ABiMS bioinformatics team, we went over the initial proof of concept and built a primary powerful and extensible analysis environment. Through first results, we obtained fund from Auvergne and Bretagne district councils (calls for projects LIFEGRID3 and BIOGENOUEST – CORSAIRE dispositive) for the next two years (2014-2015) with concrete application in fields of Nutrition and Marine Environment. The aims of future developments will be to increase metabolomic analysis workflow possibilities and to federate action to other scientist groups involved in Metabolomics through a Galaxy environment. In the near future, the workflow will be shared through the Galaxy ToolShed. Finally, we intend to develop our links, related to the national research infrastructures (METABOHUB and EMBRC-France) to build a reliable metabolomic workflow infrastructure supported by the IFB in the future.

## Acknowledgements

This work was supported by ANR program ‘Investissements d’Avenir’ (METABOHUB, EMBRC-France), and by Auvergne and Bretagne district councils.

## References

[1] J.K. Nicholson, J.C. Lindon and E. Holmes, 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29(11):1181-1189, 1999.

[2] W. Dunn, Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology*, 5:011001, 2008.

[3] C.A. Smith, E.J. Want, G.C. Tong, R. Abagyan, and G. Siuzdak, XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779-787, 2006.

[4] M. Katajamaa and M. Oresic, Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics*, 6:179, 2005.

[5] R. Tautenhahn, G.J. Patti, D. Rinehart and G. Siuzdak, XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Analytical Chemistry*, 84(11):5035-5039, 2012.

[6] J. Xia, N. Psychogios, N. Young and D.S. Wishart, MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res*, 37(Web Server issue):W652-60, 2009.

[7] J. Goecks, A. Nekrutenko, J. Taylor and The Galaxy Team, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol, 11(8):R86, 2010.

[8] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko and J. Taylor, Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, 89:19.10.1-19.10.21, 2010.

[9] 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. *Genome Research*, 15(10):1451-5, 2005.

[10] D. Blankenberg, A. Gordon, G. Von Kuster, N. Coraor, J. Taylor, A. Nekrutenko and Galaxy Team, Manipulation of FASTQ data with Galaxy. *Bioinformatics*, 26(14):1783-1785, 2010.

[11] F.M. Van Der Kloet, I. Bobeldijk, E.R. Verheij, R.H. Jellema, Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. *J Proteome Res*, 8(11):5132-41, 2009.

[12] A.J. Saldanha, Java Treeview--extensible visualization of microarray data. *Bioinformatics*, 20(17):3246-3248, 2004.

# Visualization of time-series data in the context of metabolic networks with Systrip Software

Jonathan DUBOIS[1], Ludovic COTTRET[2], Amine GHOZLANE[3], David AUBER[1], Frédéric BRINGAUD[4], Patricia THÉBAULT[1,5], Fabien JOURDAN[6] and Romain BOURQUI[1]

[1] LaBRI, CNRS UMR5800, Université de Bordeaux, 351, cours de la Libration F-33405, Talence cedex, France
{jonathan.dubois, david.auber, patricia.thebault, romain.bourqui}@labri.fr
[2] LIPM, UMR INRA-CNRS, BP 52627, 31326 Castanet Tolosan Cedex, France
ludovic.cottret@toulouse.inra.fr
[3] DSIMB, INSERM UMR-S 665, Université Paris 7, INTS, 6, rue Alexandre Cabanel, 75739 Paris Cedex 15, France
amine.ghozlane@inserm.fr
[4] RMSB, UMR 5536 CNRS, Université Bordeaux Segalen, 146, rue Léo Saignat, 33076 Bordeaux cedex, France
bringaud@rmsb.u-bordeaux2.fr
[5] CGFB - CBiB, Université Victor Segalen Bordeaux 2, 146, rue Léo Saignat, 33076 Bordeaux , France
[6] INRA UMR1331 Toxalim, 180, chemin de Tournefeuille, BP 3 31931 Toulouse CEDEX, France
fabien.jourdan@toulouse.inra.fr

**Abstract** *Technological advances in biology lead to a profusion of quantitative data, raising analytical challenges. Information visualization is particularly well suited to address these difficulties as it helps to interactively move through the different levels of analysis and to simultaneously investigate data with different point of views. In this article we present, Systrip, a visual environment for the analysis of time-series data in the context of metabolic networks. Systrip gathers bioinformatics and graph theoretical algorithms that can be assembled in different ways to help bioinformaticians/biologists in their visual mining process.*

**Keywords** metabolic network, time-series data, contextual visualization.

## 1 Introduction

These last years, a technological gap has been filled in molecular biology with the development of high sequencing technologies. While it took years to sequence the genome of a single organism (e.g. fourteen for the human genome), it can now be done within a few days (and soon a few hours). It opens a new area in terms of biological investigation since the so called *omics* technologies can generate a large amount of high throughput data for the newly sequenced organisms. Global biological studies are oftenly designed to investigate the adaptation of the organism metabolism to an environmental stress (e.g. drug treatment). Samples are collected all along the adaptation process. The output data will be, for each biological entity (e.g. metabolite), a vector of numerical values corresponding to measures made at different time points. These kinds of data are called time series data. However, most general biological network visualization software only superimposes biological measurements on nodes for one time point (*e.g. Cytoscape* [11], *Pathway Tools* [10], or *BiologicalNetworks* [2]). Other tools, such as *VANTED* [8] and *VisANT* [7], support the visualization of the whole experimental data by embedding an expression profile plot, a bar chart or line chart inside the nodes.

In this article we briefly present Systrip, a visual environment for the analysis of time-series data in the context of metabolic networks (for more details please refer to [4]). The originality of Systrip is to combine bioinformatics methods (e.g. scope detection [6], chokepoint detection [12]) with well-known theoretical graph ones (e.g strahler, strength). Systrip also supports exploration awareness as it enables the biologists to manually create views on the data at each important filtration step of their analysis process. We demonstrated some measurable benefits when using Systrip to monitor the metabolism of *Trypanosoma brucei*, which is the causative agent of human African trypanosomiasis (sleeping sickness).

## 2 Software System

Systrip is developed in C++ (see http://tulip.labri.fr/TulipDrupal/?q=systrip for executable and source code) and is based on Tulip [1]. Tulip is an information visualization framework dedicated to the analysis and visualization of relational data. It aims at providing to the developer a complete library supporting the design of interactive information visualization applications.

### 2.1 Main Features

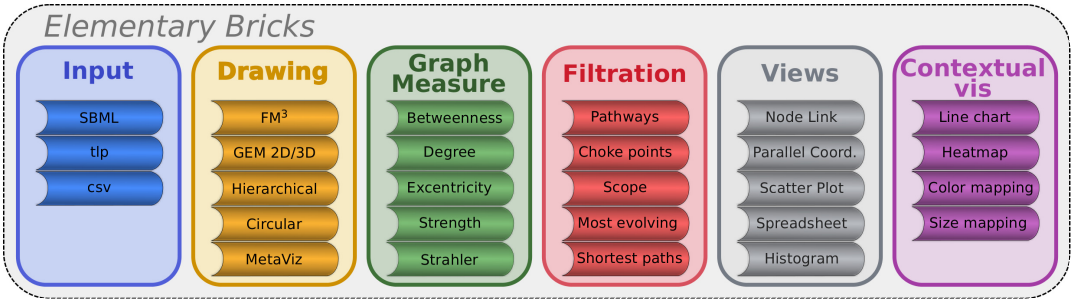



**Figure 1.** Systrip provides an environment gathering input facilities, graph drawing, graph measure and filtration algorithms, and supports several visualization techniques.

Primary aim of Systrip is to support visual analysis of time-series data in the context of the complete metabolic network of an organism. To achieve that goal, Systrip provides 5 classes of functionalities (that we call *bricks*): input, drawing, graph measure, filtration and views (see Fig. 1). These elementary bricks can also be classified into network related (e.g. betweenness centrality or excentricity), bioinformatics related (e.g. scope selection [6] or chokepoint selection [12]) and multi-dimensional data related (e.g. parallel coordinates or scatterplot). By combining several simple but also complex bricks in a meaningful pipeline, the user can define his own specifications and therefore build an image that answers the original biological question. Unlike tools like Cytoscape, Systrip disconnects the original data from the focussed subnetworks and related views. As the subnetworks are fully disconnected from the original network, the user can easily but also safely modify them (e.g. adding/removing elements, computing graph measures or applying metabolic-dedicated algorithms) without altering the original data. That feature is particularly suited to the top-down analysis process used in biology and bioinformatics. It also enables to perform simultaneously several analytics pipelines on the same data by applying different combinations of elementary bricks. Fig. 2 shows a screenshot of Systrip. In this figure, one can see on the left side a panel containing (top-left) the elements and pathways of the metabolic network and (bottom-left) all subnetworks created by the user (also containing a list of elements and pathways). The middle part of the tool contains the different views created by the user on the metabolic network, the time-series data but also database queries results. Since all the data are related to elements of the metabolic network, these views (except the information views) can be synchronized. We support a linking and brushing technique that allows to highlight selected elements in all views.

*Visualizing metabolic network:* In Fig. 2, one can see two node-link diagram views on the *Trypanosoma brucei* metabolic network. In this representation, a node is either a metabolite or a reaction, and, a metabolite and a reaction are linked if and only if the metabolite is consumed or produced by the reaction. Systrip supports various drawing algorithms (hierarchical [1], 2D/3D force-directed [5] and MetaViz [3] layout algorithms), bioinformatics algorithms (scope selection [6], chokepoints selection [12]) but also graph theoretical measures (betweenness centrality, eccentricity, strahler, strength).

*Visualizing time-series data:* Systrip supports different visualization methods to mine time-series -or any stamped- data. When focussing on individual time points, the user may use a spreadsheet or an histogram to visualize the values associated to the monitored elements (e.g. concentration or expression levels). If the user is interested in comparing values over several time points, Sytrip offers a scatterplot matrix view or parallel coordinates one. While scatterplot supports pairwise comparison and therefore evaluation of correlations

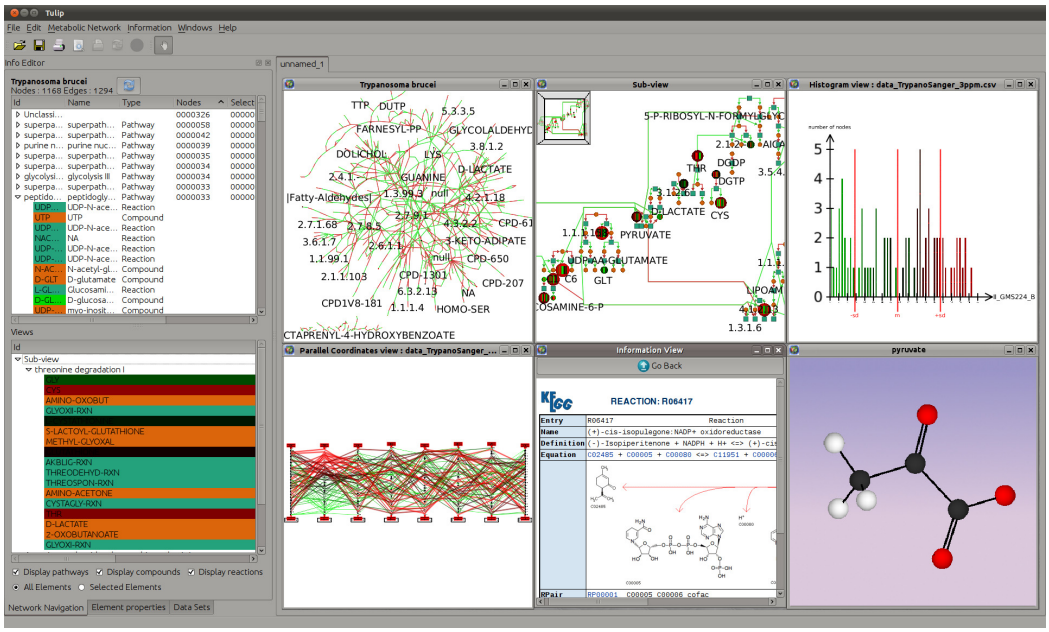


**Figure 2.** Screenshot of Systrip software. (Left) A panel containing (top-left) the elements and pathways of the metabolic network and (bottom-left) all subnetworks created by the user (also containing a list of elements and pathways); (Right) Representations of the metabolic network and subnetworks, time-series data and also database queries results.

between two dimensions (or time points), parallel coordinates allows to compare simultaneously elements over an arbitrary number of dimensions.

*Contextual visualization:* The user can also visualize time-series data in the context of the metabolic network. In that case, one can again distinguish two tasks in the analysis (i) the comparison of values of the elements at a single time point (ii) their evolutions over the entire experiment. For the comparison of values of the elements at a single time point, Systrip supports both node color and node size mapping to encode the node values for that particular time point (in Fig. 2, size and color mappings are used). To show the evolutions over the entire experiments, Systrip supports the visualization of the whole time-series of each element by rendering a heatmap or curvemap [9] glyph inside it. Another option is to navigate through the time points and to follow the evolution of the node sizes and/or colors. To preserve the user mental map, a smooth animation is performed during each transition between a time point to the next one by interpolating the node sizes and/or colors.

*Visualizing database queries:* The elements of the metabolic network are related to a large range of databases. In Systrip, the user can query both KEGG and PublicHouse databases. To visualize database results, Systrip offers an information view embedding a web browser (*e.g.* Fig. 2 shows *NADP+ oxidoreductase* as described in KEGG) as well as a 3D molecular geometry visualization obtained from the OpenBabel library (Fig. 2 shows the 3D geometry of *pyruvate*).

## 2.2 Simple vs Advanced user interface

Systrip integrates many features that can be complex to correctly set for a non-expert in graph visualization and analysis. According to the bioinformatician feedbacks, we decided to set a simple user interface as default, and, on demand, to display advanced options. The simple interface contains the most useful algorithms, i.e. mostly algorithms dedicated to biological networks. To make the interface even simpler, the parameters are automatically set to default values (these values were either set according to their biological meaning or according to empirical tests). The advanced user interface is designed for expert users and offers more functionalities than the simple one. First of all, an “Algorithms” menu, containing Tulip plug-ins installed on the computer, is added in the menu bar. Using that advanced interface, the user can set manually the needed parameters when applying an algorithm. The left panel (see Fig. 2) is also enriched: a new tab to manage the views and interaction tools settings is inserted and a configuration widget is added in the “Data Sets” tab.

## 3 Conclusion and future work

We presented Systrip, a visualization environment dedicated to metabolic networks. Selection and extraction of subnetworks based on the classification of the nodes in reference to metabolic pathways or to biological features such as biosynthetic capability have been facilitated. One of the main contributions of Systrip for the metabolic network visualization is the powerful interface to support contextual visualization of time-series data. The entire time series can be visualized as heatmap and curvemap glyphs while single time points can be mapped to node size and/or color. Additional ways to explore time-series data are proposed: spreadsheets, histograms, scatterplots and parallel coordinates. Each way to explore data can be used at the same time and mapped onto the metabolic network and eventually the extracted subnetworks. Selections on any of the representations are reflected in the other ones.

An interesting direction is to offer the ability of loading metabolic networks directly from public databases. Moreover, an actual challenge in system biology is to integrate various sources of omics data with the objective to improve the realistic significance of bioinformatics analyses. Regarding Systrip, the integration of new methods dedicated to the visualization of specific and various omics data would be on purpose and may be carried by extending the actual well defined framework.

## Acknowledgements

This work was partially done under the Systryp project, supported by the ANR (France) and BBSRC (UK); and under the EVIDEN project (ANR 2010 JCJC 0201 01), supported by the ANR (France).

## References

[1] D. Auber. Tulip : A huge graph visualisation framework. In P. Mutzel and M. Jünger, editors, *Graph Drawing Softwares*, Mathematics and Visualization, pages 105–126. Springer-Verlag, 2003.

[2] M. Baitaluk, M. Sedova, A. Ray, and A. Gupta. BiologicalNetworks: Visualization and analysis tool for systems biology. *Nucleic Acids Res.*, 34:W466–W471, 2006.

[3] R. Bourqui, V. Lacroix, L. Cottret, D. Auber, P. Mary, M.-F. Sagot, and F. Jourdan. Metabolic network visualization eliminating node redundance and preserving metabolic pathways. *BMC Systems Biology*, 1(29), 2007.

[4] J. Dubois, L. Cottret, A. Ghozlane, D. Auber, F. Bringaud, P. Thebault, F. Jourdan, and R. Bourqui. Systrip: a visual environment for the investigation of time-series data in the context of metabolic networks. In *Proc. of the 16th International Conference on Information Visualization (IV'12)*, 2012.

[5] S. Hachul and M. Jünger. Drawing large graphs with a potential-field-based multilevel algorithm. In *Proc. Graph Drawing 2004*, pages 285–295, 2004.

[6] T. Handorf, O. Ebenhöh, and R. Heinrich. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J Mol Evol*, 61(4):498–512, Oct 2005.

[7] Z. Hu, D. M. Ng, T. Yamada, C. Chen, S. Kawashima, J. Mellor, B. Linghu, M. Kanehisa, J. M. Stuart, and C. DeLisi. VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res.*, 35:W625–W632, 2007.

[8] B. H. Junker, C. Klukas, and F. Schreiber. VANTED: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7:109, 2006.

[9] M. Meyer, B. Wong, M. Styczynski, T. Munzner, and H. Pfister. Pathline: A Tool for Comparative Functional Genomics. *Computer Graphics Forum*, 29, 2010.

[10] S. Pailey and P. Karp. The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Research*, 34(13):3771–3778, 2006.

[11] P. Shannon, A. Markiel, O. Ozierand, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13:2498–2504, 2003.

[12] I. Yeh, T. Hanekamp, S. Tsoka, P. Karp, and R. Altman. Computational analysis of plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome Res*, 14(5):917–924, May 2004.

# Towards a Life Sciences Virtual Research Environment

## An e-Science initiative in Western France

Yvan LE BRAS[1], Aurélien ROULT[1], Cyril MONJEAUD[1], Mathieu BAHIN[1,2], Olivier QUENEZ[1], Claudia HERIVEAU[1], Anthony BRETAUDEAU[1,3], Olivier SALLOU[1] and Olivier COLLIN[1]

[1] GenOuest Core Facility, UMR6074 IRISA CNRS/INRIA/Université de Rennes1, Campus de Beaulieu, 35042, Rennes Cedex, France
Yvan.le_bras@irisa.fr

[2] EcoBio, UMR6553 CNRS Université de Rennes1, avenue Leclerc, Campus de Beaulieu, 35042, Rennes Cedex, France

[3] INRA IGEPP, UMR1349 Agrocampus-Ouest INRA Université Rennes1, domaine de la motte, 35653, Le Rheu, Cedex 35327, France

**Abstract** *Research processes in Life sciences are evolving at a rapid pace. This evolution, mainly due to technological advances, offers more powerful equipment and generalizes the digital format of research data. In the data deluge context, we need to overcome the current "datanami" and prepare for the future. In Life Sciences, we are noting a sharp increase of storage and computing needs. The current model, consisting to regularly add hardware resources on the Bio-informatics core facilities without global coordination, is no longer sustainable. Scientific data management and analysis has to be enhanced in order to offer services and developments corresponding to the new uses. Using Information and Communication Technology (ICT) as international standards and softwares (ISAtools software suite for metadata management, Galaxy web platform for data analysis and HUBzero for scientific collaboration), we propose a life sciences Virtual Research Environment (VRE) for Western France Science communities. If deployment of this kind of environment is challenging, it represents an opportunity to pave the way towards better research processes through enhanced collaboration, data management, analysis practices and resources optimization.*

**Keywords** eScience, VRE, collaboration, ICT, data management, metadata management, data analysis, data sharing, scientific computing, storage.

## Vers un Environnement Virtuel de Recherche en Sciences de la Vie

### Une initiative e-Science dans le Grand Ouest

**Résumé** *Les modalités des pratiques en recherche en Biologie évoluent à un rythme soutenu. Cette évolution, fortement couplée au progrès technologiques, permet de disposer d'appareillages plus puissants et généralise la digitalisation des données de la recherche. Dans le contexte actuel de déluge de données, nous devons faire face au "datanami" actuel et préparer les prochaines vagues. En Sciences de la Vie, nous observons une forte augmentation des besoins en calcul et stockage. Le modèle actuel, consistant à ajouter de façon régulière des ressources matérielles sur les plateformes Bio-informatique sans coordination globale, n'est plus suffisant. La gestion et l'analyse des données scientifiques doivent être améliorées afin d'offrir des services et des développements en accord avec les nouveaux usages. En utilisant les Technologies de l'Information et de la Communication (TIC) ainsi que des logiciels et standards internationaux (ISAtools suite pour la gestion des métadonnées, la plateforme web Galaxy pour l'analyse de données et HUBzero pour les aspects collaboration scientifiques), nous proposons un environnement virtuel de recherche (EVR) dédié aux communautés des sciences de la vie du grand Ouest.*

**Mots-clés** eScience, VRE, collaboration, TIC, gestion de données, gestion des métadonnées, analyse de données, partage de données, calcul scientifique, stockage.

## 1 Introduction

Le concept d'e-Science a été développé par John Taylor en 2000 alors qu'il était directeur général des conseils de recherche au Royaume-Uni. De manière synthétique, on peut considérer qu'il s'agit de la combinaison de la science et des nouvelles technologies de l'information et de la communication (TIC). L'apport principal des TIC est la possibilité de collaborer plus efficacement. Les scientifiques sont alors en mesure de coopérer en partageant les ressources de calcul, les données, les instruments, etc. Cela donne naissance à de nouveaux usages avec la possibilité de travailler en réseau, de façon collaborative et multidisciplinaire. Le caractère distribué des ressources impose qu'elles soient interconnectées par le réseau. La e-Science s'appuie donc sur une infrastructure importante : machines, réseau et applications qui constituent ce qui est appelé e-infrastructure. L'accès à une telle structure autorise les scientifiques à construire des projets ambitieux et relever des challenges auparavant inenvisageables. Nous pouvons citer par exemple : iPlant (USA), eCPC (e-Science for Cancer Prevention and Cure - Suède), eEcology (Pays Bas). Ces projets scientifiques, peuvent être enrichis par une approche de science « citoyenne » avec l'aide d'une large communauté de non-spécialistes. Le cas emblématique est celui de GalaxyZoo qui a été étendu dans le projet Zooniverse aux données spatiales, climatiques, environnementales et ethnologiques.

Les modalités et les usages de la recherche en Sciences de la Vie vivent actuellement une période de transition rapide sous l'impulsion des évolutions technologiques. Les principales améliorations concernent l'automatisation des appareillages et la production des résultats sous forme numérique. La combinaison de ces deux facteurs représente une opportunité d'accélérer les processus et d'élargir les champs de la recherche. Le caractère numérique des données facilite par exemple une réutilisation de ces dernières par la communauté. Mais elle constitue également un challenge. Les nouveaux appareillages se caractérisent en effet par une production élevée de données. Cet afflux pose alors divers problèmes que doivent affronter les chercheurs : analyse, gestion et traitement d'une grande masse de données. L'exemple le plus présent est celui de la génomique, pour laquelle le développement des nouvelles technologies de séquençage (NGS), a provoqué une rupture nette de la progression de la production de séquences avec celle des ressources de stockage et de calcul, respectivement représentées par les loi de Kryder et Moore [1]. Cela implique le fait que le modèle actuel de fonctionnement et d'évolution des plateformes bio-informatiques, à savoir l'ajout continuel de ressources matérielles pour faire face aux demandes, ne constitue pas une solution durable. Il est donc impératif de se doter des méthodes, techniques et infrastructures adéquates afin d'assurer l' exploitation optimale des données produites.

Le recourt à la simulation, ajouté à l'augmentation des données produites dans le cadre des sciences expérimentales, a provoqué une forte augmentation généralisée de la genèse de données. A travers cette production massive et le caractère digital des processus de recherche, nous sommes finalement entrés dans une science intensivement liée à la donnée et au calcul. Cette nouvelle aire représente le quatrième paradigme de l'exploration scientifique [2,3] et elle se manifeste notamment par l'émergence des disciplines X-info ou Comp-X (ex : Bio-informatique, Chemo-informatique, Astro-informatique). L'environnement accompagnant les processus de recherche est également issu d'une évolution importante, liée au développement des outils numériques. Les clusters, grilles et cloud se banalisent aujourd'hui. En sciences de la vie, nous n'échappons pas à cette évolution et cela renforce les liens existants entre le monde de la biologie et celui de l'informatique. Pour pallier aux manques de compétences de chacun de ces deux domaines vis-à-vis de l'autre, une vision tripartite Biologie / Bio-informatique / Informatique est aujourd'hui indispensable. Ainsi, l'optimisation des processus informatiques et bio-informatiques se répercute aujourd'hui directement sur le bon déroulement des projets de recherche en sciences de la vie.

A l'heure actuelle, une forte demande émane des scientifiques de tous horizons pour disposer de plus de ressources matérielles. Pourtant, plus que l'équipement, ce sont les usages sur lesquels nous devrions concentrer nos efforts. Il semble essentiel de permettre une gestion intégrée des ressources, d'ouvrir nos communautés à la mutualisation et de mettre en place des couches applicatives permettant aux scientifiques d'accéder aux ressources et de collaborer. C'est dans ce cadre que nous proposons aujourd'hui un environnement virtuel de recherche (Fig. 1), élément clé de la mise en place d'une telle approche.

Appelé VRE (Virtual Research Environment) [4], cet environnement permet de gérer l'ensemble de plus en plus complexe des tâches requises pour mener à bien un projet de recherche. Il s'agit d' un écosystème mettant à disposition des utilisateurs, via une interface web, les logiciels, les données, les ressources

systèmes et d'administration. Nous présentons ici une première initiative de mise en place d'un VRE orienté sciences de la vie à destination des scientifiques français du Grand Ouest. Il se base sur l'utilisation de solutions telles que la plateforme collaborative HUBzero, la suite logicielle ISA et le portail d'analyse de données Galaxy et permet un accès unique à la collaboration scientifique, à la gestion des métadonnées ainsi qu'à l'analyse des données de la recherche.

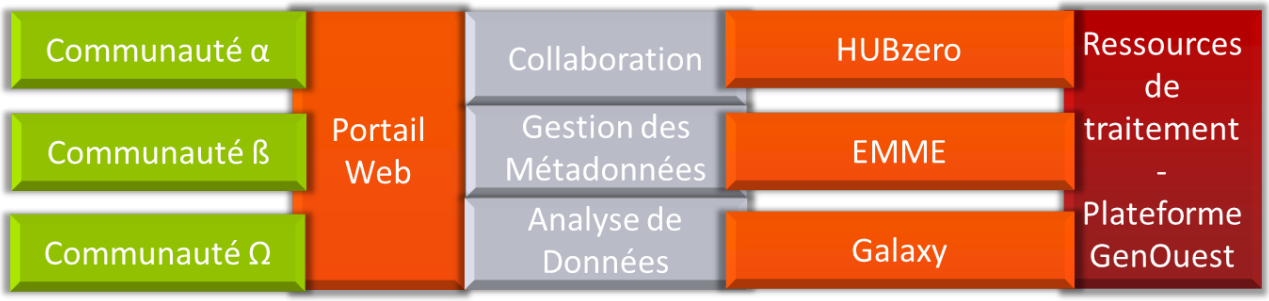



**Figure 1 : Vision synthétique d'un Environnement Virtuel de Recherche. Un VRE permet, en optimisant les processus de collaboration, de relier ses utilisateurs aux données, logiciels et ressources de traitement nécessaires à la genèse et au partage de connaissances spécifique de leurs communautés.**

## 2 Un accès unique à la collaboration et aux ressources

La notion d'accès unique est essentielle à la mise en place d'un VRE. Pour que l'accès soit possible à des utilisateurs distribués sur un territoire, il doit s'effectuer via une interface web. Pour notre environnement de recherche, nous avons opté pour le déploiement d'une solution basée sur la plateforme HUBzero [5]. Si d'autres solutions proposent des fonctionnalités similaires, comme Alfresco [6], Laboratree [7] ou MS SharePoint [8], HUBzero [5] est la seule solution développée spécialement pour le milieu de la recherche et proposant le plus de fonctionnalités [9]. Elle est de plus open-source et très évolutive.

### 2.1 Présentation de HUBzero

HUBzero [5] a été développé par l'université de Purdue, en collaboration avec le réseau NCN (Network for Computational Nanotechnology) et sponsorisé par la National Science Foundation, pour le calcul en nanotechnologie. Développé à partir du gestionnaire de contenu (CMS) Joomla!, il intègre de nombreux composants développés spécifiquement pour la collaboration scientifique. Si l'utilisation de Joomla! ne fait pas l'unanimité auprès des administrateurs systèmes, elle facilite la création et la gestion des affichages graphiques ainsi que la personnalisation des fonctionnalités à travers les nombreuses extensions que propose ce CMS. HUBzero [5] propose initialement l'incorporation d'outils de calcul et de simulation via l'environnement RAPPTURE [10]. nous avons ici préféré déporter les activités centrées sur les données via d'autres environnements, des portails, décrits plus loin, jugés mieux adaptés aux besoins des communautés visées. Nous proposons donc aux scientifiques d'utiliser notre HUB e-Biogenouest comme portail d'accès au VRE et comme plateforme de collaboration.

### 2.2 Un outil centré sur la collaboration

La plateforme e-Biogenouest offre la possibilité aux utilisateurs d'accéder à de nombreuses fonctions collaboratives (Fig. 2). A travers la gestion des utilisateurs et des groupes, le HUB eBGO permet ainsi la tenue et l'enrichissement de wikis, de forums de discussion, de blogs, de calendriers, de listes de souhaits et de partage de ressources multimédia au sein des communautés constituées. Outre l'aspect réseau social, cette appartenance à un ou plusieurs groupes permet notamment de faciliter les activités d'animation nécessaires au maintien de groupes de travail et à la diffusion d'information provenant de media sous diverses formes (images, vidéos, document pdf, …).

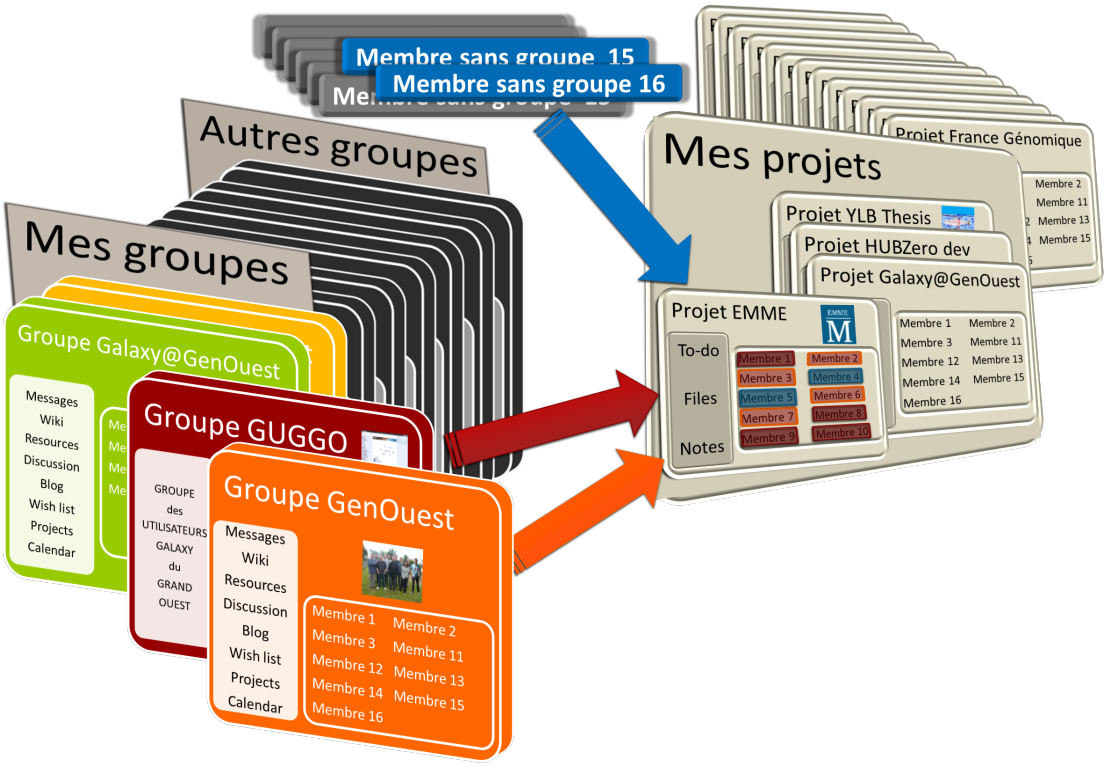



**Figure 2 : Représentation schématique de la structuration en groupes et projets au sein du HUB. Les membres enregistrés sur le HUB eBGO, peuvent créer et rejoindre des groupes et des projets. L'appartenance à un groupe facilite l'animation et les interactions (échanges de ressources et d'informations) entre membres d'une même communauté. L'appartenance à un projet facilite le partage de fichiers et permet d'avoir accès à un environnement simplifié de gestion de projet.**

Les ressources échangeables sont actuellement classées en 12 catégories : cours, séminaires, présentations sous forme de diapositives, manuels, tutoriaux, flux vidéos, téléchargements, publications, thèses, jeux de données, données accessibles par URL et ressources accessibles par URL. Le nombre et la nature de ces catégories sont modifiables. L'utilisateur peut de cette manière contribuer à la base de connaissance que représente ce HUB, étiqueter et noter les contenus.

Les membres et groupes d'eBGO peuvent également créer et/ou rejoindre des projets. Au sein de ces projets, correspondants à des projets de recherche en cours ou des projets de développement ou de test, les membres ont accès à un environnement simplifié de gestion de projet. Ils peuvent ainsi créer et affecter des tâches et les organiser en listes thématiques. Un espace d'échange de fichiers est également mis à disposition des membres du projet. Il s'agit d'une bonne solution pour partager des rapports de début, étape ou fin de projet, des comptes rendus de réunions ou encore des fichiers présentant une synthèse des analyses effectuées. Une gestion de version de ces fichiers est également proposé via l'utilisation transparente d'un dépôt basé sur Git. Enfin, un espace est réservé à la rédaction de notes en utilisant le formatage de type wiki proposé par HUBzero. Ces notes permettent par exemple de faire une synthèse de premières étapes d'analyses, en pointant vers les fichiers de synthèse ou compte rendus de réunion de la section "Files" et en proposant l'intégration de divers contenus comme des images ou des liens vers du contenus interne à eBGO. Les notes produites peuvent être liées de façon hiérarchique.

Les développeurs d'HUBzero ont récemment annoncé la mise en place dans les mois à venir d'un système plus poussé de gestion et hiérarchisation des données au sein des projets.

## 2.3 Un portail d'accès au VRE

Bien que ce HUB représente l'élément non centré sur la donnée de notre VRE, il permet le lien essentiel entre les groupes d'utilisateurs, les projets de recherche et la donnée. Il est en effet le point de départ réunissant les informations nécessaires pour accéder aux plateformes de gestion et d'analyse de données elles-mêmes connectées aux ressources de calcul. Il permet ainsi de proposer un accès direct aux outils et méthodes utilisées mais également aux données, métadonnées et références associées.

Après la phase de conception de l'étude, un projet de recherche débute par une collecte de données. Dès lors, s'enchaînent les étapes de traitement, de diffusion et d'analyse permettant de faire émerger des résultats. La gestion des données et métadonnées conditionne alors la connaissance extraite ainsi que la qualité, la réutilisation et la pérennité de l'information qu'elles contiennent.

## 3 Gestion des métadonnées

Le rôle joué par la "métadonnée" est de plus en plus important dans nos communautés. Cette information renseignant sur les données de la recherche se retrouve entre autre dans des cahiers de laboratoire, sur des tubes, dans les documents liés à la demande de financement de projet, ou bien encore, comme dans les cas des données d'imagerie, au niveau de la donnée elle-même. Elle est également représentée sous forme de données. Il apparaît aujourd'hui évident que cette information, une fois passée à un format digital, permet de faciliter les tâches liées à la gestion et à l'analyse des données de la recherche. Pourtant, les scientifiques n'accordent souvent que trop peu de temps à leurs enregistrement et structuration, tâches non obligatoires à l'obtention de résultats et à la création de connaissance.

Les récentes orientations de la recherche prennent de plus en plus conscience de l'importance de ces métadonnées. Des plans de gestion de données sont désormais demandés lors de réponses à des appels à projet. La soumission d'articles dans de nombreuses revues nécessite l'enregistrement préalable des données et métadonnées associées dans des dépôts publics internationaux. Mais ces tâches sont bien souvent fastidieuses et l'origine très hétérogène des données ne facilite pas ces étapes de gestion. Pour permettre de meilleures pratiques, il est essentiel de pouvoir proposer des outils configurables se basant sur des standards internationaux, utilisant des ontologies contrôlées et validées par les différentes communautés. C'est dans ce cadre que se place la démarche d'ISA [11].

### 3.1 ISAtools et le format ISA-Tab

La suite d'outils open source de gestion de métadonnées ISA [11] facilite l'enregistrement, la conservation et la réutilisation des données dans un contexte de conformité vis-à-vis des standards internationaux en sciences de la vie. Le système ISA [11] se base sur une hiérarchie dont le plus haut niveau correspond au contexte du projet ("Investigation") suivi par l'unité de projet de recherche ("Study") pour finir par les mesures analytiques ("Assay").

Pour enregistrer les informations relatives aux projets de recherches, le système de fichiers se base sur un format tabulé, nommé ISA-Tab. Ce format de fichier a été préféré vis-à-vis d'autres formats tel que le xml car il est facile à créer, visualiser et éditer par les chercheurs utilisant bien souvent des tableurs de type MS Excel® [12]. De plus, il existe des API permettant la manipulation de ces fichiers tabulés. Nous trouverons dans ces fichiers, des champs renseignant sur les technologies et protocoles utilisés, la référence des différents fichiers traités et générés, de la donnée brute jusqu'à la donnée normalisée et transformée, ou encore les publications générées par le projet.

La suite d'outils proposée permet alors de créer et modifier ces fichiers ISA-Tab. Elle offre également de nombreuses fonctionnalités comme la possibilité de créer ses propres configurations de mesures analytiques, en fonction du type de mesure effectuée et de technologie employée. Il est également possible de convertir un fichier de type ISA-Tab dans un format accepté pour la soumission dans des dépôts publics tels que ENA (génomique), PRIDE (protéomique) et ArrayExpress (transcriptomique). Il est enfin intéressant de noter qu'aujourd'hui, le format ISA-Tab représente également un format de soumission, comme pour le journal GigaScience [1].

### 3.2 L'application web BioInvestigation Index (Bii)

Associé au format ISA-Tab, les développeurs d'ISAtools ont également mis à disposition de la communauté scientifique un modèle pour le stockage des métadonnées ainsi qu'une interface web pour la consultation. Nommé Bii pour "BioInvestigation Index", il s'agit d'une infrastructure open source permettant de stocker et représenter les données et métadonnées issues de projets de recherche.

[1] http://www.gigasciencejournal.com/

### 3.3 Gestion de données et métadonnées au sein du VRE

Si les outils ISA ont pour but d'être exécutés en local, notre environnement propose via le portail EMME [13] (Fig. 3), un accès facilité via le web, sans authentification préalable. La génération d'une arborescence de fichiers dans les dossiers de l'utilisateur et le chargement d'une configuration personnalisée développée pour les communautés du Grand Ouest se fait ensuite de manière automatique.

A partir de ces archives ISA, notre VRE propose aux utilisateurs, via une instance Bii, de soumettre les données et métadonnées liées à leurs projets de recherche mais également d'explorer ceux de la communauté. Nous offrons également la possibilité de lier les données et métadonnées renseignées dans des archives ISA (personnelles ou accessible via notre Bii) à notre instance de la plateforme web d'analyse de données Galaxy [14, 15, 16]. L'utilisateur a ainsi la possibilité de traiter directement les archives ISA compressées. Il peut notamment récupérer des jeux de données stockés sur un serveur distant dont l'URL a été renseigné lors de la création de l'archive.

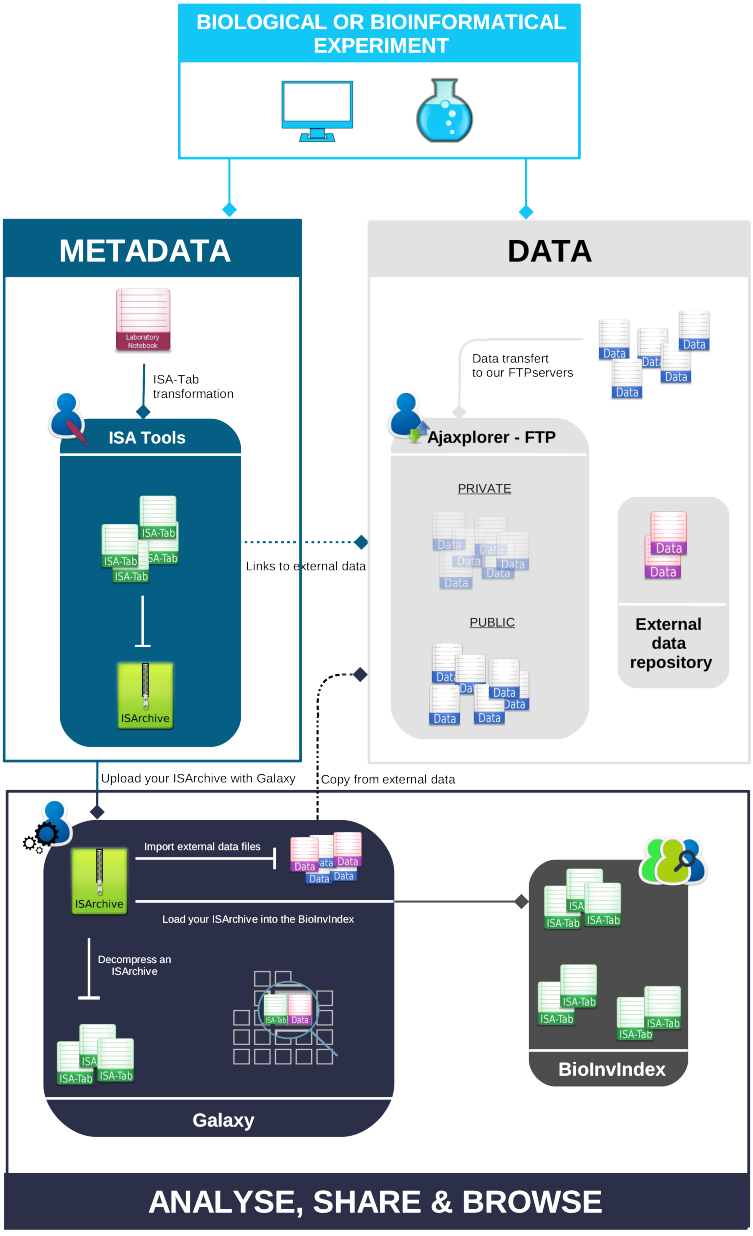



Figure 3 : Représentation schématique des différents composants du portail EMME[2]

[2] http://emme.genouest.org/

# 4 Analyse et gestion des données

Les outils d'analyses de données biologiques ne manquent pas. Que l'on s'intéresse par exemple à la génomique, à la protéomique, à la génétique des populations, à la modélisation des systèmes biologiques ou à l'écologie, c'est à chaque fois un véritable arsenal d'outils qui s'offre à nous. D'un manière générale, les outils à disposition sont soit basés sur la connaissance des langages de programmation, comme lors de l'utilisation de R [17], Python [18] ou Perl [19], soit "prêts à l'emploi", comme MS Excel® [12], GenespringGX [20], Proteome Discoverer [21], Arlequin [22] ou Orange [23]. Depuis plusieurs années, l'afflux de données, la nécessité de personnaliser, relier les outils, d'utiliser des infrastructures de calcul performantes, ajouté au prix des licences des solutions "tout-en-un" ont poussé de plus en plus de biologistes à se tourner vers la première catégorie. Les conséquences de cette évolution du métier de chercheur sont nombreuses et non négligeables. L'outil informatique en général n'étant pas bien connu et maitrisé, notamment par les communautés des sciences de la vie, le coût de cette diversification est important. Nous pouvons citer notamment un investissement non négligeable en temps de la part des scientifiques lors des processus de formation. Les scripts élaborés sont bien souvent peu optimisés en terme d'utilisation de ressources matérielles et de temps d'exécution. Ils sont généralement mal documentés, ne permettant que difficilement leur partage. Enfin, à travers ces activités d'écriture de scripts, nous assistons à un fort doublonnage des processus de recherche. D'une manière générale, les outils actuels rendent difficile le suivi des traitements appliqués aux données, le partage de fichiers ou de visualisation ainsi que l'optimisation de l'utilisation des ressources matérielles de stockage, de calcul ou d'échange. Pour remédier à ces nombreux problèmes, des solutions de plateformes d'analyses basées sur l'utilisation du web ont vu le jour. Si nous proposons depuis plusieurs années un portail basé sur Mobyle [24], l'orientation NGS de nombreux de projets de la communauté du Grand Ouest représentait une opportunité de déployer et tester une instance de Galaxy [14, 15, 16].

## 4.1 La plateforme web d'analyse de données Galaxy

Galaxy [14, 15, 16] est une plateforme d'analyse de données proposant, comme Mobyle [24], d'utiliser des ressources de calculs performantes, des scripts édités sous différents langages (Python, Perl, R, Shell, …) via une interface web et sans que l'utilisateur ait de connaissances en programmation. Sans rentrer dans le détail, cette solution permet entre autre aux utilisateurs de visualiser et traiter des fichiers de plusieurs millions de lignes. Elle intègre un grand nombre d'outils permettant la manipulation de fichiers, l'analyse de séquences ou l'application de nombreuses opérations sur les données. Au sein de cet environnement d'analyse, le scientifique peut ainsi enchaîner les outils, garder la trace des tâches appliquées aux données sous forme d'historiques et créer des workflows reprenant l'enchaînement de ces tâches. Il peut également visualiser des données génomiques en utilisant le composant de visualisation Trackster intégré à Galaxy. Tous ces éléments peuvent être privés, publics ou être partagés avec certains utilisateurs. Les scripts et les données ainsi que l'accès aux applications étant centralisés au niveau d'un serveur unique, leur utilisation et partage peuvent être réalisés en limitant les effets de doublonnage.

## 4.2 Galaxy au sein du VRE

Si Galaxy [14, 15, 16] est à l'origine clairement orienté vers la génomique, l'architecture de cette plateforme permet l'intégration aisée d'outils de diverses origines. Ainsi, nous avons développé et adapté de nombreux outils pour permettre d'ouvrir les possibilités de cet environnement aux diverses communautés scientifiques d'intérêt. Nous mettons désormais à disposition des utilisateurs des outils permettant entre autre de traiter des données de protéomique, phylogénie, génétique des populations et génétique quantitative. Nous interfaçons également des outils développés par la communauté rendant plus facile leur accès et test. Dans ce VRE, les fonctionnalités de partage proposés par Galaxy prennent tout leur sens puisqu'elles permettent un accès rapide et sans utilisation importante des infrastructures réseaux pour accéder aux données, historiques, workflows et visualisations via l'utilisation de leurs URL, par exemple intégrées dans une page d'un forum ou d'un wiki de HUBzero [5]. Actuellement, l'accès à notre instance de Galaxy (http://galaxy.genouest.org/) s'effectue suite à une authentification auprès du LDAP de la plateforme GenOuest. Cette étape nécessite

donc la création préalable d'un compte. La liste des outils proposés et les worklows et pages publiques de notre instance sont toutefois disponibles sur eBGO[3].

## 5 Conclusion et retours d'expérience

Le schéma proposé ici est un schéma général (Fig. 4) qui peut répondre à une importante portion des besoins numériques des chercheurs en Sciences de la vie. Les choix fait en terme de solutions logicielles étaient principalement guidés par la recherche d'outils matures, flexibles, déjà adoptés par la communauté scientifique et s'appuyant sur des standards internationaux. Ce n'est qu'à partir de 2011, date de la naissance du projet, que l'ensemble des outils du VRE présenté ici ont intégrés chacun de ses aspects. Il est à noter que les technologies considérées individuellement sont accessibles et font déjà partie de l'arsenal de bon nombre de scientifiques. C'est dans l'intégration et l'adoption de ces technologies que se trouvent souvent les verrous.

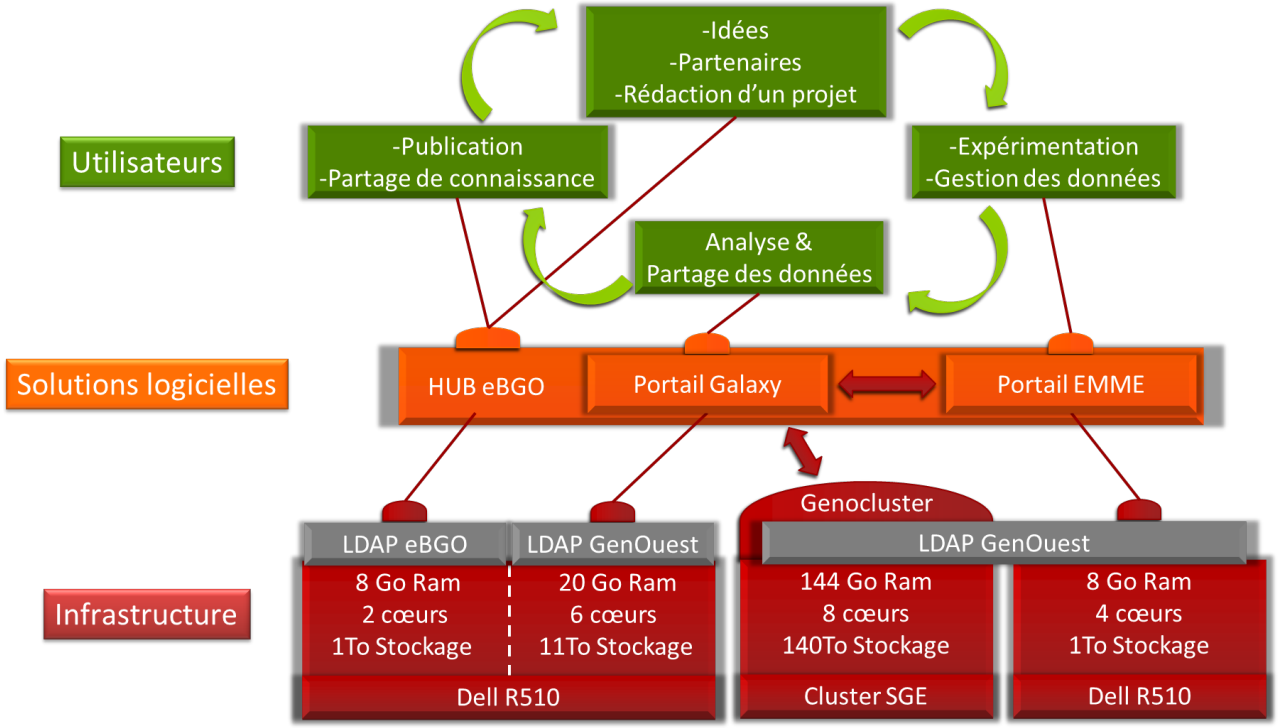



**Figure 2 : Schéma du VRE déployé, des processus d'un projet de recherche en sciences de la vie à l'infrastructure matérielle via les solutions logicielles.**

La mise en place et l'administration de cet environnement a nécessité l'implication quotidienne d'un administrateur système et de 2 développeurs / administrateurs principaux épaulés par 5 autres développeurs et sous le guidage d'un coordinateur. Nous avons la chance dans le Grand Ouest de posséder des liens forts avec les différentes communautés scientifiques notamment via le réseau Biogenouest. Ceci facilite les échanges et la mise en place de phases de tests ainsi que les étapes de développement, et donc d'adaptation de notre environnement à leurs besoins. La flexibilité est l'une des particularités essentielles d'un VRE. Elle permettra ainsi l'intégration d'autres solutions de gestion et d'analyse de données en fonction des besoins des communautés non ciblées actuellement par notre environnement.

Si notre VRE représente un environnement relativement complet, nous pourrions déplorer l'absence actuelle d'outils de gestion avancée des ressources bibliographiques. Bien que HUBzero [5] permet là aussi de créer, répertorier, annoter et partager des ressources documentaires, un lien avec une solution de gestion de références bibliographiques de type Zotero [25], CiteUlike [26] ou Mendeley [27] est à explorer.

A l'heure de la mise en place de grandes infrastructures nationales, telles que l'IFB (Institut Français de Bio-informatique) pour la Bio-informatique, FBI (France Bio-Imaging) ou FLI (France Life Imaging) pour la

[3] https://www.e-biogenouest.org/einfrastructure/data/analysis

Bio-imagerie, il parait essentiel de mettre en place des structures de niveau intermédiaire, afin de pouvoir permettre un portage optimisé des projets scientifiques vers les ressources adéquates. Il n'y a, à notre connaissance, pas d'autres initiatives de ce type en sciences de la vie au niveau français, et ce premier test pourra certainement alimenter les réflexions autour de la structuration de la recherche dans les autres régions. Un mésocentre orienté sciences de la vie propose en effet une proximité indispensable avec les scientifiques porteurs des projets, tout en représentant une interface vers les structures nationales mais également les organisations de plus grande portée comme les réseaux européens ELIXIR ou EuroBioImaging.

Associé à des mésocentres, la mise en place de VRE visant les communautés régionales, devrait permettre une optimisation de l'utilisation des ressources matérielles et du déroulement des processus de recherche. Vu les modes de financement actuels, trop ponctuels et fluctuant, une stratégie coordonnée inter-organismes s'avère nécessaire à la mise en place d'une telle infrastructure. Un intérêt du concept e-Science réside alors dans le fait qu'il autorise plusieurs approches stratégiques. Il est ainsi envisageable de créer une e-infrastructure permettant la réalisation de développements locaux tout en étant en mesure de basculer les projets sur des infrastructures de niveau national ou international. Il est également possible de se reposer exclusivement ou en combinaison, sur des ressources externes, académiques et privées.

La stratégie mise en place propose d'optimiser les infrastructures matérielles existantes tout en permettant son enrichissement avec les données. L'émergence de telles infrastructures de données, devenues un élément clé de la e-Science, permettront l'accélération des avancées scientifiques. La réussite d'une telle entreprise repose sur l'acquisition des compétences nécessaires pour développer les couches applicatives requises. L'élément clé est donc l'investissement dans des compétences humaines, notamment à travers la mutualisation de la matière grise.

## Acknowledgements

This work was supported by Brittany and Pays de la Loire Regions, CNRS, Biogenouest.

## References

[1] Kahn, On the future of genomic data. *Science,* 331:728-9,2011.

[2] G. Bell, T. Hey, and A. Szalay, Beyond the Data Deluge. *Science*, 323 (5919): 1297−1298, 2009

[3] T. Hey, S. Tansley, K. Tolle, The fourth paradigm, data-intensive scientific discovery. *Redmond, Washington: Microsoft Research*, 2009

[4] L. Candela, Data use − Virtual Research Environments. Technological & Organisational Aspects of a Global Research Data Infrastructure - A view from the expert http://www.grdi2020.eu/Pages/SelectedDocument.aspx.

[5] M. McLennan, R. Kennell, HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering. *Comput Sci Eng*, 12:48-53, 2010.

[6] Alfresco: Open Source Enterprise Content Management System (CMS). Available: http://www.alfresco.com.

[7] Selican Technologies, Inc. Available: http://selican.com.

[8] Collaboration Software for the Enterprise−SharePoint 2010. Available: http://sharepoint.microsoft.com.

[9] A. E. Berman, W. K. Barnett, S. D. Mooney, Collaborative software for traditional and translational research. *Hum. Genomics*, 6:21, 2012.

[10] W. Qiao, M. McLennan, R. Kennell, D.S. Ebert, G. Klimeck, Hub-based simulation and graphics hardware accelerated visualization for nanotechnology applications. *IEEE Trans Vis Comput Graph*., 12:1061-8, 2006.

[11] P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 26:2354-2356, 2010.

[12] Microsoft, Microsoft Excel. Redmond, Washington: Microsoft, 2010.

[13] C. Monjeaud, Y. Le Bras, O. Collin, EMME : an Experimental Metadata Management Environment. *JOBIM 2013,* poster session.

[14] J. Goecks, A. Nekrutenko, 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.

[15] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, J. Taylor, Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology,* Chapter 19:Unit 19.10.1-21, 2007.

[16] B. Giardine, C. Riemer, RC. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, WJ. Kent, A. Nekrutenko, Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451-5, 2005.

[17] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2011 http://www.R-project.org/.

[18] G. van Rossum et al., Python Language Website. Available: http://www.python.org/.

[19] L. Wall, Perl, the first postmodern computer language. Available: http://www.perl.org/.

[20] Agilent Technologies, GeneSpring. Available: http://www.agilent.com.

[21] Thermo Scientific, Proteome Discoverer Software. Available: http://www.thermoscientific.com/.

[22] Excoffier, L. G. Laval, and S. Schneider, Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 1:47-50, 2005.

[23] Orange. Available: http://www.ailab.si/orange

[24] B. Néron, H. Ménager, C. Maufrais, N. Joly, J. Maupetit, S. Letort, S. Carrere, P. Tuffery, and C. Letondal, Mobyle: a new full web bioinformatics framework. *Bioinformatics* 25: 3005-3011, 2009.

[25] Anon, Zotero: The Next-Generation Research Tool. 2008. Available: http://www.zotero.org/.

[26] K. Emamy, RG. Cameron, Citeulike: A researcher's social bookmarking service. *Ariadne* 51, 2007. Available: http://www.ariadne.ac.uk/issue51/emamy-cameron/.

[27] Anon, Mendeley–Manage and Share Research Papers–Discover Research Data. 2008. Available: http://www.mendeley.com.

# Session 3C : Génomique d'association et analyse de variants

# Network-guided multi-locus association mapping with graph cuts

Chloé-Agathe AZENCOTT[1], Dominik GRIMM[1], Mahito SUGIYAMA[1], Yoshinobu KAWAHARA[2] and Karsten BORGWARDT[1,3]

[1] Machine Learning and Computational Biology Research Group, Max Planck Institute for Developmental Biology & Max Planck Institute for Intelligent Systems Spemannstr. 38, 72076 Tübingen, Allemagne
{cazencott,dominik.grimm,mahito.sugiyama,karsten.borgwardt}@tuebingen.mpg.de
[2] The Institute of Scientific and Industrial Research (ISIR) Osaka University 8-1 Mihogaoka, Ibaraki-shi, Osaka 567-0047 Japon
kawahara@ar.sanken.osaka-u.ac.jp
[3] Zentrum für Bioinformatik, Eberhard Karls Universität Tübingen, 72076 Tübingen, Allemagne

**Abstract** *As an increasing number of genome-wide association studies reveal the limitations of attempting to explain phenotypic heritability by single genetic loci, there is growing interest for associating complex phenotypes with* sets of genetic loci. *While several methods for multi-locus mapping have been proposed, it is often unclear how to relate the detected loci to the growing knowledge about gene pathways and networks. The few methods that take biological pathways or networks into account are either restricted to investigating a limited number of predetermined sets of loci, or do not scale to genome-wide settings.*

*We present SConES, a new efficient method to discover sets of genetic loci that are maximally associated with a phenotype, while being connected on an underlying network. Our approach is based on a minimum cut reformulation of the problem of selecting features under sparsity and structural constraints, which can be solved exactly and rapidly.*

*SConES outperforms state-of-the-art competitors in terms of runtime, scales to hundreds of thousands of genetic loci, and exhibits higher power in detecting causal SNPs in simulation studies than existing methods. On flowering time phenotypes and genotypes from* Arabidopsis thaliana*, SConES detects loci that enable accurate phenotype prediction and that are supported by the literature.*

**Keywords** GWAS, association mapping, multi-locus, biological networks, graph cuts

## Détection d'association multi-locus guidée par des réseaux biologiques au moyen de coupes de graphes

**Résumé** *Alors que les limites d'études d'associations pangénomiques visant à expliquer un trait phénotypiques uniquement par l'effet de loci génétiques uniques sont de plus en plus apparentes, l'intérêt de la communauté pour les approches utilisant des* ensembles de loci *se fait croissant. Bien que de nombreuses méthodes multi-locus aient déjà été proposées, leur interprétabilité comme leur pouvoir de détection restent limités, et il est difficile de faire le lien entre ces études et les connaissances biologiques à notre disposition sous formes de voies de signalisations ou de réseaux. En effet, les rares méthodes prenant ces derniers en compte sont limitées à l'étude d'un nombre restreint d'ensembles prédéterminés de loci et sont difficiles à appliquer au large nombre de variables que l'on observe dans un cadre génome entier.*

*Nous présentons ici SConES, un algorithme nouveau et efficace pour la découverte d'ensembles de loci génétiques qui sont simultanément maximalement associés avec un phénotype et connectés sur un réseau sous-jacent. Notre approche repose sur une reformulation coupe-minimale du problème de la sélection de variables sous contraintes de parcimonie et de connectivité qui permet une solution rapide et exacte.*

*SConES est plus rapide que l'état de l'art, est aisément applicable à des centaines de milliers de loci génétiques, et est plus performant que ses compétiteurs pour détecter les SNPs réellement causatifs dans des simulations. Dans le cas de phénotypes liés au temps de floraison d'*Arabidopsis thaliana*, SConES détecte des loci qui permettent une bonne prédiction phétnotypique et sont corroborés par la litérature scientifique.*

**Mots-clés** études d'association, loci multiples, réseaux biologiques, coupes de graphes

## 1 Introduction

Twin and family/pedigree studies make it possible to estimate the heritability of observed traits, that is to say the amount of their variability that can be attributed to genetic differences. In the past few years, genome-wide association studies (GWAS), in which several hundreds of thousands to millions of single nucleotide polymorphisms (SNPs) are assayed in up to thousands of individuals, have made it possible to identify hundreds of genetic variants associated with complex phenotypes [33]. Unfortunately, while studies associating single SNPs with phenotypic outcomes have become standard, they often fail to explain much of the heritability of complex traits [21]. Investigating the joint effects of multiple loci by mapping sets of genetic variants to the phenotype has the potential to help explain part of this missing heritability [22]. While efficient multiple linear regression approaches [6, 31, 26] make the detection of such multivariate associations possible, they often remain limited in power and hard to interpret. Incorporating biological knowledge into these approaches could help boosting their power and interpretability. However, current methods are limited to predefining a reasonable number of candidate sets to investigate [5, 8, 32], for instance by relying on gene pathways. They consequently run the risk of missing biologically relevant loci that have not been included in the candidate sets. This risk is further increased by the incomplete state of our current biological knowledge.

For this reason, our goal here is to use prior knowledge in a more flexible way. We propose to use a biological network, defined between SNPs, to guide a multi-locus mapping approach that is both efficient to compute and biologically meaningful: *We aim to find a set of SNPs that (a) are maximally associated with a given phenotype and (b) tend to be connected in a given biological network. In addition, this set must be computed efficiently on genome-wide data*. In this paper we assume an additive model to characterize multi-locus association. Let us stress that the method must scale to networks of hundreds of thousands or millions of nodes, and that approaches developed to analyze gene networks containing hundreds of nodes, such as that of [24], [7] or [18] do generally not apply.

Our task is a feature selection problem in a graph-structured feature space, where the features are the SNPs and the selection criterion should be related to their association with the phenotype considered. Several approaches have already been developed to address such problems.

The overlapping group Lasso [12, 19] is a sparse linear model designed to select features that belong to the union of a small number of predefined groups. If a graph over the features is given, defining those groups as all pairs of features connected by an edge or as all linear subgraphs of a given size yields the so-called graph Lasso. A similar approach is taken by [11]: their structured sparsity penalty encourages selecting a small number of base blocks, where blocks are sets of features defined so as to match the structure of the problem. In the case of a graph-induced structure, blocks are defined as small connected components of that graph. As shown in [20], the overlapping group Lasso mentioned above is a relaxation of this binary problem. Furthermore, [18] propose a network-constrained version of the Lasso that imposes the type of graph connectivity we deem desirable. However, none of these approaches easily scales to graphs over hundreds of thousands or millions of nodes.

In the case of directed acyclic graphs, [20] propose a minimum flow formulation that makes it possible to use for groups (or blocks) the set of all paths of the network. Unfortunately, the generalization to undirected graphs with cycles, such as the SNP networks we consider, requires to randomly assign directions to edges and prune those in cycles without any biological justification. Although this can work reasonably well in practice,this is akin to artificially removing more than half of the network connections without any biological justification.

In what follows, we propose a minimum cut formulation of the network-guided SNP selection problem in Section 2 and evaluate the performance of our solution both in simulations and on actual *Arabidopsis thaliana* data in Section 3.

## 2 Methods

Let $n$ be the number of SNPs and $m$ the number of individuals. The SNP-SNP network is described by its adjacency matrix $\boldsymbol{W}$ of size $n \times n$. A number of statistics based on covariance matrices, such as HSIC [9]

or SKAT [32], can be used to compute a measure of dependence $\boldsymbol{c} \in \mathbb{R}^n$ between each single SNP and the phenotype. Under the common assumption that the joint effect of several SNPs is additive (which corresponds to using linear kernels in those methods), $\boldsymbol{c}$ is such that the association between a group of SNPs and the phenotype can be quantified as the sum of the scores of the SNPs belonging to this group. In other words, given an indicator vector $\boldsymbol{f} \in \{0,1\}^n$ such that, for any $p \in \{1, \cdots, n\}$, $f_p$ is set to 1 if the $p$-th SNP is selected and 0 otherwise, the score of the selected SNPs is given by $Q(\boldsymbol{f}) = \sum_{p=1}^{n} c_p f_p = \boldsymbol{c}^\top \boldsymbol{f}$.

We want to find the indicator vector $\boldsymbol{f} \in \{0,1\}^n$ that maximizes the score $Q(\boldsymbol{f})$ while ensuring that the solution is made of connected components of the SNP network. Rather than searching through all subgraphs of a given network, we reward the selection of adjacent features through graph regularization. As it is also desirable for biological interpretation, and to avoid selecting large number of SNPs in linkage disequilibrium, that the selected sub-networks are small in size, we reward sparse solutions. The first requirement can be addressed by means of a smoothness regularizer on the network [28, 2], while the second one can be enforced with an $l_0$ constraint:

$$\underset{\boldsymbol{f}\in\{0,1\}^n}{\arg\max} \underbrace{\boldsymbol{c}^\top \boldsymbol{f}}_{\text{association}} - \underbrace{\lambda\, \boldsymbol{f}^\top \boldsymbol{L} \boldsymbol{f}}_{\text{connectivity}} - \underbrace{\eta\ ||\boldsymbol{f}||_0}_{\text{sparsity}} \tag{1}$$

where $\boldsymbol{L}$ is the Laplacian of the SNP network. $\boldsymbol{L}$ is defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{D}$ is the diagonal matrix where $\boldsymbol{D}_{p,p}$ is the degree of node $p$. Note that here, we directly minimize the number of non-zero entries in $f$ and do not require the proxy of an $l_1$ constraint to achieve sparsity (of course in the case of binary indicators, $l_1$ and $l_0$ norms are equivalent). $\lambda$ and $\eta$ are positive parameters that control the importance of the connectedness of selected features and the sparsity regularizer, respectively.

PROPOSITION 2.1. *Given a graph $\mathcal{G}$ of adjacency matrix $\boldsymbol{W}$, solving the graph-regularized feature selection problem formalized in Eq. 1 is equivalent to finding an s/t min-cut on the graphwhose vertices are that of $\mathcal{G}$, augmented by two additional nodes $s$ and $t$, and whose edges are given by the adjacency matrix $\boldsymbol{A}$, where $\boldsymbol{A}_{p,q} = \lambda \boldsymbol{W}_{p,q}$ for $1 \leq p, q \leq n$ and*

$$\boldsymbol{A}_{s,p} = \begin{cases} c_p - \eta \text{ if } c_p > \eta \\ 0 \quad \text{otherwise} \end{cases} \text{ and } \boldsymbol{A}_{t,p} = \begin{cases} \eta - c_p \text{ if } c_p < \eta \\ 0 \quad \text{otherwise} \end{cases} \quad (p = 1, \ldots, n).$$

It is therefore possible to use maximal flow algorithms to efficiently optimize the objective function defined in Equation (1) and select a small number of connected SNPs maximally associated with a phenotype. In our implementation, we use the Boykov-Kolmogorov algorithm [4]. Although its worst case complexity is in $\mathcal{O}(n^2 n_E n_C)$, where $n_E$ is the number of edges of the graph and $n_C$ the size of the minimum cut, it performs much better in practice, particularly when the graph is sparse. We refer to this method as SConES, for Selecting CONnected Explanatory SNPs.

## 3 Results

We evaluate the ability of SConES to detect networks of trait-associated SNPs on simulated datasets and on datasets from an association mapping study in *Arabidopsis thaliana*.

### 3.1 Experimental Settings

In all of our experiments, the association term $\boldsymbol{c}$ of SConES is derived from Linear SKAT [32], which makes it possible to correct for covariates (and therefore population structure). SKAT has been devised to address rare variants associations problems by grouping SNPs to achieve statistical significance, but can equally be applied to common variants.

*Comparison partners* We compare SConES to the following algorithms: a univariate linear regression (where the SNPs selected are those with a Bonferroni-corrected $p$-value $\leq 0.05$); a Lasso regression; the network-constrained Lasso (ncLasso) of [18]; and groupLasso and graphLasso [12].

*Setting the parameters* All methods considered, except for the univariate linear regression, have parameters (e.g. $\lambda$ and $\eta$ in the case of SConES) that need to be optimized. In our experiments, we run 10-fold cross-validation grid-search experiments over ranges of values of the parameters: 7 values of $\lambda$ and $\eta$ each for SConES and ncLasso, and 7 values of the parameter $\lambda$ for the Lasso and the non-overlapping group Lasso (ranging from $10^{-3}$ to $10^3$). We then pick as optimal the parameters leading to the most stable selection, and report as finally selected the features selected in all folds. More specifically, we define stability according to a consistency index similar to that of [15]: The consistency index between two feature sets $S$ and $S'$ is defined as $I_C(S, S') = \frac{n|S \cap S'| - |S||S'|}{n \min(|S|,|S'|) - |S||S'|}$. For an experiment with $k$ folds, the consistency is computed as the average of the $k(k-1)/2$ pairwise consistencies between sets of selected features.

### 3.2 Simulations

To assess the performance of our methods, we simulate phenotypes for $m = 500$ real *Arabidopsis thaliana* genotypes (214 051 SNPs), chosen at random among those made available by [10]. Restricting ourselves to $1,000$ randomly picked SNPs with minor allele frequency larger than $10\%$, we pick 20 of the SNPs to be causal, and generate phenotypes $y_i = w^\top g_i + \epsilon$, where both the support weights $w$ and the noise $\epsilon$ are normally distributed.

We consider the following networks:

- *Genomic sequence network* (GS): SNPs adjacent on the genomic sequence are linked together. In this setting we aim at recovering sub-sequences of the genomic sequence that correlate with the phenotype.
- *Gene membership network* (GM): SNPs are connected as in the sequence network described above; in addition, SNPs within 20kb of the same gene are linked together as well. For groupLasso, GM groups are defined so that all SNPs near the same gene belong to the same group.
- *Gene interaction network* (GI): SNPs are connected as in the gene membership network described above. In addition, SNPs within 20kb of two genes connected on the TAIR [29] protein-protein interaction network are linked together (or grouped together for groupLasso).

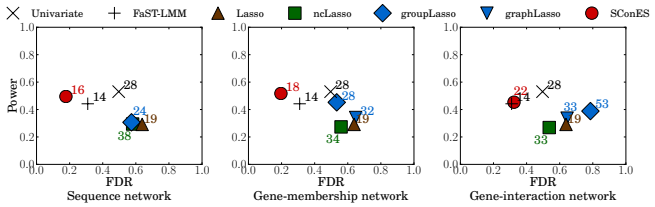
On various simulation scenarios, we evaluate the different methods in terms of power (fraction of true causal SNPs selected) and false discovery rate (FDR, fraction of selected SNPs that are not causal). As SConES returns a binary feature selection rather than a feature ranking, it is not possible to draw FDR curves or compare powers at same FDR as is often done when evaluating such methods. Nevertheless, we observe that SConES is generally better than its comparison partners according to both criteria, as can be seen on Fig. 1. The only exception is that groupLasso is superior when the groups it uses perfectly match the causal structure (an unlikely event in practice).

We further observe that SConES is robust to missing edges: its performance hardly varied when removing up to $12\%$ of the edges at random. Moreover, it is less sensitive to the quality of the structure used than its comparison Lasso approaches. Nevertheless, its performance, like that of all other network-regularized approaches, decreases when the causal SNPs are more spread out in the network, and is strongly negatively affected when the network is entirely inappropriate.

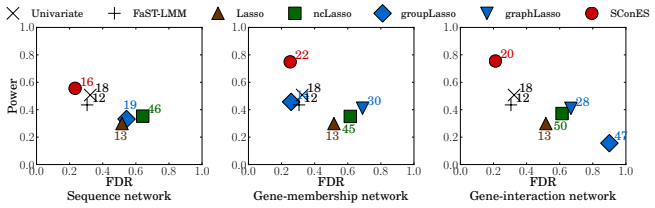
Finally, ncLasso is both slower and less performant than SConES. This indicates that solving the feature selection problem we pose directly, rather than its relaxed version, allows for better recovery of true causal features.

### 3.3 *Arabidopsis* Flowering Time Phenotypes

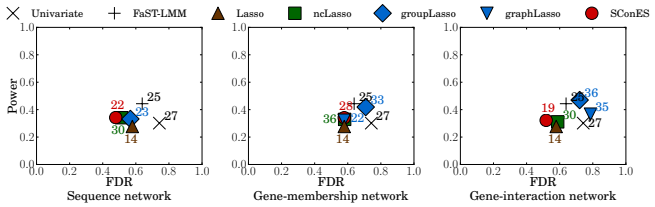
We then apply our method to a large collection of 17 *A. thaliana* flowering times phenotypes from [3] (up to 194 individuals, 214 051 SNPs). The groups and networks are again derived from the TAIR protein-protein interaction data. We filter out SNPs with a minor allele frequency lower than $10\%$, as is typical in *A.*




(a) Scenario (1): The true causal SNPs belong to the same genomic segment




(b) Scenario (2): The true causal SNPs are near the same gene




(c) Scenario (3): The true causal SNPs are near any of five interacting genes

**Figure 1.** Power and false discovery rate (FDR) of SConES, compared to state-of-the-art Lasso algorithms and a baseline univariate linear regression, in three different data simulation scenarios. Best methods are closest to the upper-left corner. Numbers denote the number of SNPs selected by the method.

*thaliana* GWAS studies. We use the first principal components of the genotypic data as covariates to correct for population structure [25].

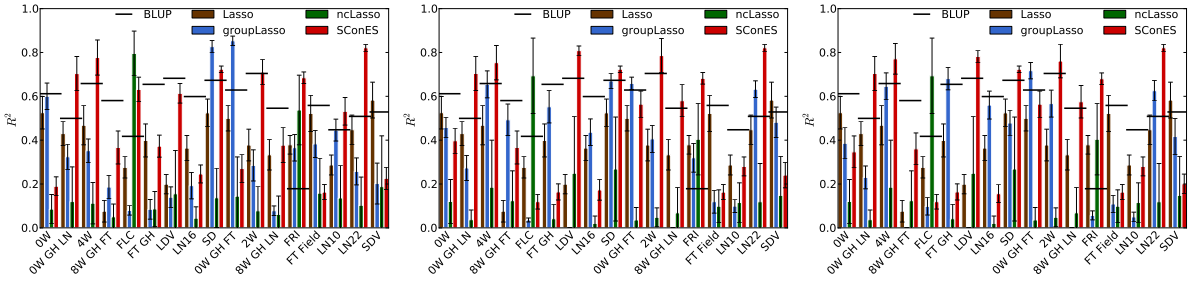
As graphLasso does not scale to datasets such as ours with more than $2.10^5$ SNPs, we exclude it from our experiments. We run Lasso, ncLasso, groupLasso and SConES on the flowering time phenotypes as described in Section 3.1. However, for many of the phenotypes, the Lasso approaches select large number of SNPs (more than $10\,000$), which makes the results hard to interprete. Using cross-validated predictivity, as is generally done for Lasso, still does not entirely solve this issue, particularly for large group sizes. We therefore filter out solutions containing more than $1\%$ of the total number of SNPs before using consistency to select the optimal parameters.

To evaluate the quality of the SNPs selected, we perform ridge regression on each phenotype in a cross-validation scheme that uses only the selected SNPs and report its average Pearson's squared correlation coefficient in Fig. 2. The superiority of groupLasso in that respect is to be expected, as predicitivity is directly optimized by the regression. Indeed, the features selected by groupLasso+GS achieve higher predictivity than SConES+GS on most phenotypes. Nevertheless, the features selected by SConES+GM are at least as predictive as those selected by groupLasso+GM in two thirds of the phenotypes; the picture is the same for SConES+GI, whose features are on average more predictive than those of groupLasso+GI.

We also record how many distinct flowering time candidate genes listed by [27] are retrieved on average by the various methods. A gene is considered retrieved if the method selects a SNP near it. Our results are shown in Table 1. Methods retrieving a large fraction of SNPs near candidate genes do not necessarily retrieve the largest number of distinct candidate genes. Good predictive power, as shown in Fig. 2, however, seems to correlate with the number of distinct candidate genes selected by an algorithm, not with the percentage of selected SNPs near candidate genes. groupLasso+GI has the highest fraction of candidate gene SNPs among all methods, but detects only three distinct candidate genes in average. This is probably due to groupLasso selecting entire genes or gene pairs; if groupLasso detects a candidate gene, it will pick most of the SNPs near that gene.

| | LinReg | Lasso | groupLasso | | | ncLasso | | | SConES | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | GS | GM | GI | GS | GM | GI | GS | GM | GI |
| #SNPs | 5 | 86 | 153 | 611 | 546 | 684 | 608 | 608 | 729 | 546 | 496 |
| near candidate genes | 9% | 9% | 10% | 9% | 20% | 4% | 6% | 6% | 18% | 8% | 7% |
| candidate genes hit | 0 | 4 | 4 | 1 | 3 | 5 | 5 | 5 | 12 | 15 | 12 |

**Table 1.** Summary statistics, averaged over the *Arabidopsis thaliana* flowering time phenotypes: average total number of selected SNPs ("#SNPs"), average proportion of selected SNPs near candidate genes ("near candidate genes") and average number of different candidate genes recovered ("candidate genes hit") . "GS": Genomic sequence network. "GM": Gene membership network. "GI": Gene interaction network.




(a) Genomic sequence networks (b) Gene membership networks (c) Gene interaction networks

**Figure 2.** Cross-validated predictivity (measured as Pearson's squared correlation coefficient between actual phenotype and phenotype predicted by a ridge-regression over the selected SNPs) of SConES compared to that of Lasso, groupLasso, and ncLasso.

## 4 Discussion and Conclusions

This article presented SConES, a novel approach to multi-locus mapping that selects SNPs that tend to be connected in a given biological network without restricting the search to predefined sets of loci. As the optimization problem of SConES can be solved by maximum flow, our solution is computationally efficient and scales to whole genome data. Our experiments show that our method is one to two orders of magnitude faster than the state-of-the-art Lasso-based comparison partners, and can therefore easily scale to hundreds of thousands of SNPs. In simulations, SConES is better at leveraging the structure of the biological network to recover causal SNPs. On real GWAS data from *Arabidopsis thaliana*, the predictive ability of the features selected by SConES is superior to that of groupLasso on two of the three network types we consider. When using more biological information (gene membership and gene information), SConES tends to recover more distinct explanatory genes than groupLasso, which in turns leads to better phenotypic prediction.

SConES is less vulnerable to ill-defined networks than its comparison partners, and is robust to missing edges, which is particularly desirable in the light of the current noisiness and incompleteness of biological networks. Our results on the GS network actually indicate that graphLasso, using pairs of network edges as groups, may achieve the same flexibility as SConES; unfortunately it is too computationally demanding to be run on the most informative networks. Replacing the Laplacian by a random-walk based matrix, so as to differently treat disconnected SNPs that are closeby in the networks from those that are far apart, has the potential to further increase missing edge robustness and will be considered in future work.

We currently derive the SNP networks from neighborhood along the genome sequence, closeness to a same gene, or proximity to interacting proteins. Refining those networks and exploring other types of networks as well as understanding the effects of their topology and density is one of our next projects.

Let us note that while we do not explicitly consider linkage disequilibrium, the $l_0$ sparsity constraint of SConES should enforce that when several correlated SNPs are associated with a phenotype, a single one of them is picked. On the other hand, if SConES is given a genomic sequence network such as the one we describe, the graph smoothness constraint will encourage nearby SNPs to be selected together, leading to the selection of sub sequences that are likely to be haplotype blocks. Such a network should therefore only be used when the goal of the experiment is to detect consecutive sequences of associated SNPs.

For now SConES considers an additive model between genetic loci. Future work includes taking pairwise multiplicative effects into account. Replacing the association term in Equation (1) by a sum over pairs of SNPs rather than over individual SNPs results in a maximum flow problem over a fully connected network of SNPs, which cannot be solved straightforwardly, if only because the resulting adjacency matrix is too large to fit in memory on a regular computer. It might be possible, however, to leverage some of the techniques used for two-locus GWAS [1, 14] to help solve this problem.

An important extension of SConES is to devise a way to evaluate the statistical significance of the set of selected SNPs. Regularized feature selection approaches such as SConES or its Lasso comparison partners do not lend themselves well to the computation of $p$-values. Permutation tests could be an option, but the number of permutations to run is difficult to evaluate, as is that of hypotheses tested. Another possibility would be to implement the multiple-sample splitting approach proposed by [23]. However, the loss of power from performing selection on only subsets of the samples is too large, given the sizes of current genomic data sets, to make this feasible. Therefore evaluating statistical significance and controlling false discovery rates of Lasso and SConES approaches alike remain a challenge for the future.

Note that SConES can also be applied to the detection of shared networks of markers between multiple phenotypes. Further exciting research topics include applying it to larger data sets from human disease consortia.

## Acknowledgments

The authors would like to thank Recep Colak, Barbara Rakitsch and Nino Shervashidze for fruitful discussions. C.A. is funded by an Alexander von Humboldt fellowship.

## References

[1] P. Achlioptas, B. Schölkopf, and K. Borgwardt. Two-locus association mapping in subquadratic time. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 726–734, New York, NY, USA, 2011. ACM.

[2] R. K. Ando and T. Zhang. Learning on graph with laplacian regularization. In *Advances in Neural Information Processing Systems 19*, 2007.

[3] S. Atwell et al. Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, 2010.

[4] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1124 –1137, Sept. 2004.

[5] R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.*, 86(1):6–22, Jan. 2010.

[6] S. Cho et al. Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Ann. Hum. Genet.*, 74(5):416–428, 2010.

[7] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3(140), 2007.

[8] B. L. Fridley and J. M. Biernacka. Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur J Hum Genet*, 2011.

[9] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, pages 63–77. Springer-Verlag, 2005.

[10] M. W. Horton et al. Genome-wide patterns of genetic variation in worldwide arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*, 44(2):212–216, 2012.

[11] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 417–424, New York, NY, USA, 2009. ACM.

[12] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 433–440, New York, NY, USA, 2009. ACM.

[13] B. Jie, D. Zhang, C.-Y. Wee, and D. Shen. Structural feature selection for connectivity network-based MCI diagnosis. In P.-T. Yap et al., editors, *Multimodal Brain Image Analysis*, volume 7509 of *Lecture Notes in Computer Science*, pages 175–184. Springer Berlin / Heidelberg, 2012.

[14] T. Kam-Thong et al. GLIDE: GPU-Based Linear Regression for Detection of Epistasis. *Hum Hered*, 73:220–236, 2012.

[15] L. I. Kuncheva. A stability index for feature selection. In *Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications*. ACTA Press, 2007.

[16] B. Le Saux and H. Bunke. Feature selection for graph-based image classifiers. In J. Marques, N. Perez de la Blanca, and P. Pina, editors, *Pattern Recognition and Image Analysis*, volume 3523 of *Lecture Notes in Computer Science*, pages 147–154. Springer Berlin / Heidelberg, 2005.

[17] H. F. Lee and D. R. Dooly. Algorithms for the constrained maximum-weight connected graph problem. *Naval Research Logistics*, 43:985–1008, 1996.

[18] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[19] J. Liu, J. Huang, S. Ma, and K. Wang. Incorporating group correlations in genome-wide association studies using smoothed group lasso. *Biostat*, 2012.

[20] J. Mairal and B. Yu. Path coding penalties for directed acyclic graphs. In *Proceedings of the 4th NIPS Workshop on Optimization for Machine Learning (OPT'11)*, 2011.

[21] T. A. Manolio et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

[22] J. Marchini, P. Donnelly, and L. R. Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 37(4):413–417, 2005.

[23] N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

[24] Ş. Nacu, R. Critchley-Thorne, P. Lee, and S. Holmes. Gene expression network analysis and applications to immunology. *Bioinformatics*, 23(7):850–858, 2007.

[25] A. L. Price et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–909, 2006.

[26] B. Rakitsch, C. Lippert, O. Stegle, and K. Borgwardt. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 2012.

[27] V. Segura et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet*, 44(7):825–830, 2012.

[28] A. Smola and R. Kondor. Kernels and regularization on graphs. In B. Schölkopf and M. Warmuth, editors, *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 144–158. Springer Berlin / Heidelberg, 2003.

[29] The Arabidopsis Information Resource. TAIR Protein-Protein Interaction, 2012. http://www.arabidopsis.org/portals/proteome/proteinInteract.jsp.

[30] K. Tsuda. Graph classification methods in chemoinformatics. In H. H.-S. Lu, B. Schölkopf, and H. Zhao, editors, *Handbook of Statistical Bioinformatics*, Springer Handbooks of Computational Statistics, pages 335–351. Springer Berlin Heidelberg, 2011.

[31] D. Wang, K. Eskridge, and J. Crossa. Identifying qtls and epistasis in structured plant populations using adaptive mixed lasso. *Journal of Agricultural, Biological, and Environmental Statistics*, 16:170–184, 2011.

[32] M. C. Wu et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, 89(1):82–93, 2011.

[33] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. USA*, 109(4):1193–1198, Jan. 2012.

# Incorporating linkage disequilibrium blocks in Genome-Wide Association Studies

Alia DEHMAN[1], Christophe AMBROISE[2] and Pierre NEUVIAL[2]

Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071 – USC INRA
{alia.dehman, christophe.ambroise, pierre.neuvial}@genopole.cnrs.fr

**Abstract** *In genome-wide association studies, we are interested in finding genetic markers that are significantly associated with a phenotype of interest. Whole-genome single nucleotide polymorphism (SNP) data are collected for many thousands of SNP markers, leading to high-dimensional regression problems where the number of predictors greatly exceeds the number of observations. Moreover, these predictors are highly dependent, in particular due to linkage disequilibrium (LD).*

*We propose a two-step approach that explicitly takes advantage of the grouping structure induced by LD. In the first step, we infer LD blocks by performing a clustering of LD estimates with an adjacency constraint. In the second step, we perform Group Lasso regression on the inferred LD blocks.*

*We argue that it is relevant to assess performance both at the scale of individual SNPs and at the scale of LD blocks. We investigate the efficiency of this approach compared to state-of-the art penalized regression methods (Lasso and Elastic-Net) at these two scales. Our numerical experiments show that the proposed approach not only activates the groups containing the associated SNPs but is also as precise as the penalized models as for selecting individual associated predictors.*

**Keywords** Genome-wide association studies, penalized regression, Group Lasso, linkage disequilibrium.

## 1 Introduction

With recent advances in hight-throughput genotyping technology, genome-wide association studies (GWAS) have become an important tool for identifying genetic markers underlying a variation in a given phenotype. In GWAS, it is expected that only a subset of SNPs is significantly associated with the phenotype. Therefore, the SNP selection problem can be formulated as a variable selection problem in sparse and high-dimensional settings. SNPs can also be highly correlated due to the phenomenon of linkage disequilibrium (LD) and a natural group structure can thus be considered among the variables.

The Lasso [6] is an efficient multivariate variable selection method in high-dimensional problems. However, in presence of a group structure on the predictors with high pairwise correlations among them, the Lasso tends to select only one variable from each group and discard the others. Therefore, the "group version" of the Lasso, the Group Lasso [10], was introduced in order to account for this structure in penalized regression model. Like other group-adapted regression methods, such as structured elastic-net [5] , group-MCP [2], group-SCAD [7] or the group SMCP [4], the Group Lasso involves a structured penalty allowing sparse group regression.

As SNPs of a LD block can be correlated regardless of the phenotype, and as causal SNPs can in turn be correlated to non causal ones, we argue that it may make more sense to look for *SNP groups* (that is, LD blocks) that are significantly associated with the phenotype, rather than looking for *individual SNPs*. From a biological point of view, this is particularly relevant in a context where "causal SNP" (or, more generally, causal loci) need not be observed: it is possible that the observed data only contain SNPs that are in LD with causal ones. From a statistical perspective, the distinction between SNP-level and LD block-level associations is related to an identifiability issue: assuming that causal SNPs are observed, is their association to the phenotype strong

enough so that they can be distinguished from indirect associations between SNPs in strong LD with causal ones ?

In this paper, we propose a two-step method consisting on inferring LD blocks using a spatially-constrained hierarchical agglomerative algorithm before applying the Group Lasso regression model. This approach is compared do state-of-the-art penalized regression approaches used in high-dimensional problems, for which prior group structure information is ignored (Lasso) or incorporated less directly (Elastic-Net). Competing methods are evaluated in terms of their ability to retrieve *groups of SNPs* associated to the phenotype, and to retrieve *individual SNPs* associated to the phenotype.

The rest of the paper is organized as follows: in Section 2 we describe the proposed two-step approach. The competing mehods and the evaluation are described in Section 3. Results of our simulation studies are presented in Section 4. Finally, Section 5 provides a summary and discusses some related issues.

## 2 A Two-Step Approach Taking the Group Structure Into Account

We consider the problem of predicting a continuous response $\mathbf{y} \in \mathbb{R}^n$ from covariates $\mathbf{X} \in \mathbb{R}^{n \times p}$. For $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$, $\mathbf{X}_{i\cdot}$ is a $p$-dimensional vector of covariates for observation $i$ and $\mathbf{X}_{\cdot j}$ is a $n$-dimensional vector of observations for covariate $j$. For each $i \in \{1, \ldots, n\}$, we assume that $\mathbf{X}_{i\cdot}$ has a block structure with $G$ blocks of sizes $p_1, \ldots p_G$, with $\sum_{g=1}^{G} p_g = p$. Thus $\mathbf{X}_i = (\mathbf{X}_i^1, \ldots, \mathbf{X}_i^G)$ with each $\mathbf{X}_i^g \in \mathbb{R}^{p_g}, g = 1, \ldots, G$. We note $\beta_g$ the coefficient vector corresponding to the $g^{\text{th}}$ group. We propose a two-step method consisting in inferring the LD blocks using only the genotype data, and then performing a Group Lasso regression on the inferred blocks.

### 2.1 Inference of Blocks from Genotypes

For this algorithm's first step of inferring groups, only the genotype data are used. The LD measure $D'$ is calculated from the genotype matrix [3] to obtain a $p \times p$ matrix of pairwise $D'$ measures. Then, the Fisher transformation is applied to the LD matrix to obtain quantities that are approximately normally distributed. Finally, we perform a constrained hierarchical clustering of the LD matrix, as now described.

Our clustering procedure is based on the one of the most widely used methods of cluster analysis, the Ward's incremental sum of squares method [8]. The general goal of sum of squares clustering is to minimize the total within-cluster dispersion for $G$ groups around $G$ centroids. If we denote by $D_k$ the sum of squares (or dispersion) for cluster $k$, then the increase of dispersion after merging clusters $k$ and $l$ is $I_{kl} = D_{kl} - D_k - D_l$. The standard agglomerative hierarchical approach is to start with $p$ clusters of size 1, and to sucessively merge the two clusters $k$ and $l$ which yield the smallest increase in dispersion $I_{kl}$, until only 1 cluster of size $p$ remains. Our constrained clustering is a simple modification that takes advantage of the fact that LD matrix can be modeled as block-diagonal: at each step of the agglomerative process, we only merge clusters that are *adjacent on the genome*. By construction, this constrained clustering is much faster than the standard hierarchical clustering. The desired number of blocks currently has to be set by the user.

For the numerical experiments reported below, we have used the R package `rioja` [1] which implements this constrained hierarchical clustering. We note that this implementation requires a $p \times p$ matrix of similarities to be computed. This can be quite problematic in the context of GWAS where $p$ can be as large as $10^5$ or $10^6$. In order to overcome this limitation, it is possible to implement an algorithm where LD measures are not passed as a matrix but calculated on the fly.

### 2.2 Selection of Blocks Associated with Phenotype

Once LD blocks have been identified, we use the Group Lasso [10] to identify blocks associated with the phenotype. Well adapted to group-structured variables, the Group Lasso estimator is defined as:

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} (||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} ||\boldsymbol{\beta_g}||_2),$$

where $||.||_2$ is the Euclidean norm and $\lambda$ is a penalty parameter. The Group Lasso has the specificity of being a group selection method: by construction, the estimated coefficients within a group will either all be zero or all nonzero.

## 3 Performance Evaluation

### 3.1 Simulation Study

Our simulation setting is adapted from the model used in Wu et. al.[9]. We set $n = 200$ and $p = 512$, with 9 groups of sizes $(2, 2, 4, 8, 16, 32, 64, 128, 256)$. The ordering of the groups is drawn at random for each simulation. If $j \neq j'$ are in the same group, $cov(\mathbf{X}_{.j}, \mathbf{X}_{.j'}) = \rho$ else $cov(\mathbf{X}_{.j}, \mathbf{X}_{.j'}) = 0$. For all $j \in \{1, \ldots, p\}, \mathbf{X}_{.j}$ is generated from a p-dimensional multivariate normal distribution whose covariance matrix is a block diagonal matrix. Then, we set $X_{ij}$ to $0, 1$ or $2$ according to whether $X_{ij} < -c$, $-c < X_{ij} < c$ or $X_{ij} > c$ with $c$ the first quartile of a standard normal distribution. Finally, the first two SNPs of groups of sizes $2, 2, 4, 8$ are chosen to be associated with the phenotype. The strength of the association is calibrated by the coefficient of determination $R^2$ of the model. The parameters of the simulation are therefore $R^2$ and $\rho$.

### 3.2 ROC-Based Evaluation

Our performance assessment aims at evaluating the ability of our proposed method to distinguish true signals from noise. As the association study is block-oriented, some definitions of "true signal" need to be specified. We define a *causal SNP* as a SNP that is simulated with a non-zero regression parameter. We also define a *block-associated SNP* as a predictor that is not directly associated with the phenotype but simulated in the same block that a causal SNP, and then can be highly correlated with it. As explained in Section 1, we are interested in two types of evaluations: a *block-level evaluation*, to assess how well a given method retrieves block-associated SNPs, and a *SNP-level evaluation*, to assess how well a given method retrieves causal SNPs.

Performance is evaluated using receiver operator characteristics (ROC) curves. For a given threshold, we plot the true positive rate (TPR), which is the fraction of true positives out of the positives versus the false positive rate (FPR), which is the fraction of false positives out of the negatives. To plot the ROC curves, we first evaluate, for each method, the TPR and FPR for a grid of underlying regularization parameter values and for each simulation. Then, we aggregate the curves at fixed parameter i.e, we calculate average TPR and FPR across all simulation runs for each underlying parameter value.

### 3.3 Competing Approaches Based on Penalized Regression

The proposed approach is compared to two state-of-the art competitors that do not explicitly take the block-structure information into account: Lasso [6] and Elastic-Net [11]. The estimators of Lasso and Elastic-Net, respectively noted $\beta_{lasso}$ and $\beta_{EN}$ are defined as:

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||$$

$$\hat{\boldsymbol{\beta}}_{EN} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda_1||\boldsymbol{\beta}|| + \lambda_2||\boldsymbol{\beta}||_2^2$$

with $\lambda, \lambda_1$ and $\lambda_2$ three regularization parameters. Thanks to the $l_1$ penalty, the Lasso model encourages sparsity by setting many regression coefficients for irrelevant SNPs to exactly zero. However, the Lasso does not incorporate any information on the group structure induced by LD blocks. Like the Lasso, the Elastic-Net simultaneously performs both automatic variable selection and continuous shrinkage. Unlike the Lasso, the Elastic-Net includes a ridge ($\ell_2$) penalty which tends to select groups of correlated variables. Therefore, the Elastic-Net incorporates some prior information regarding the block structure of the data. However, unlike the proposed method, it does not take advantage of the fact that blocks are adjacent along the genome.

## 4 Results

We have performed a comprehensive simulation study, where the correlation coefficient $\rho \in \{0, 0.1, 0.2, 0.3, 0.5, 0.8\}$ and the determination coefficient $R^2 \in \{0.1, 0.2, 0.3, 0.5, 0.8\}$. We summarize below the obtained results for $R^2 = 0.2$ only: indeed, we believe that this is a low but (unfortunately) realistic value of $R^2$ in GWAS studies. The results obtained for larger values of $R^2$ were similar, in the sense that the ranking of the methods for a given $\rho$ was the same, although the performance of each particular method is obviously an increasing function of $R^2$. Each of the ROC curves has been obtained by averaging $B = 500$ simulation runs.

### 4.1 Influence of LD

In order to evaluate the influence of LD on the performance of the methods, we assessed the different approaches on data sets with different degrees of correlation $\rho$ among variables. We set the number of inferred clusters for the Group Lasso to 9 groups, that is, the (oracle) number of groups actually used for simulations. The results are summarized in FIG. 1 for $\rho \in \{0, 0.1, 0.2\}$. For larger values of $\rho$ the results are not shown because they were quite similar to the case $\rho = 0.2$, except that the performance of the Lasso deteriorates for $\rho \geq 0.5$. In our experiments, the LD within block is of the same order of magnitude as $\rho$; therefore, we see the case $\rho = 0.2$ as the most realistic.

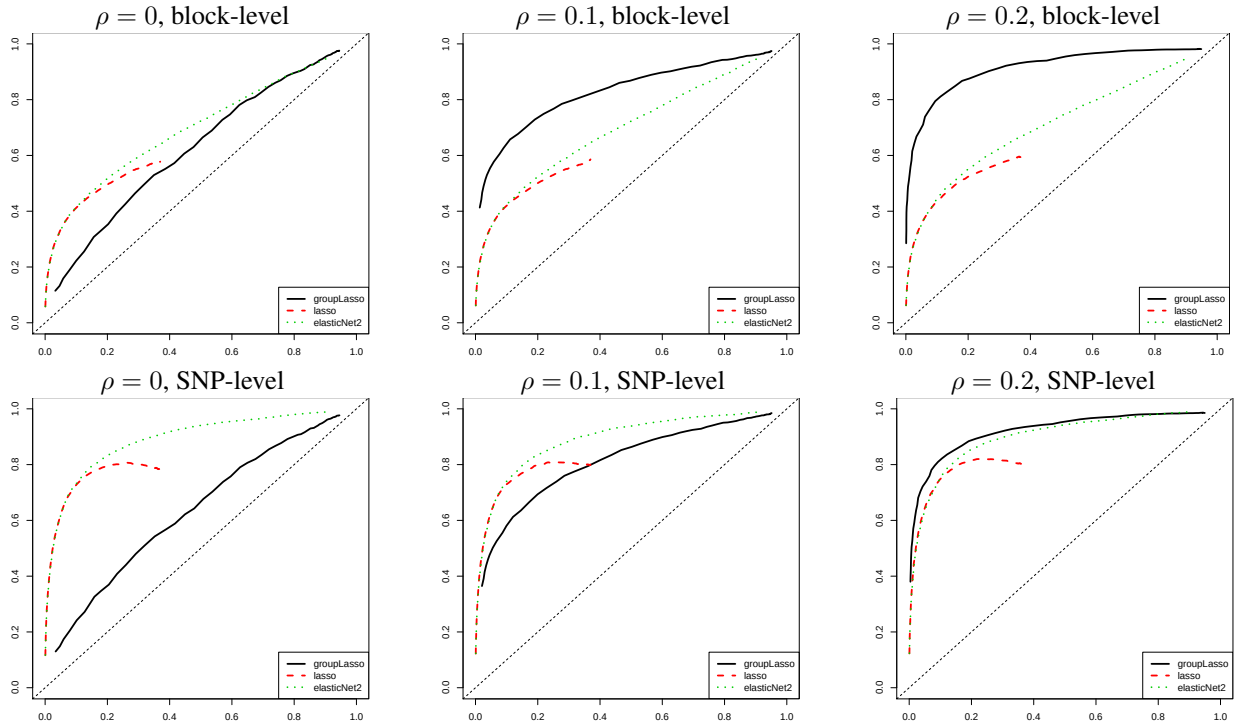



**Figure 1.** ROC curves (TPR in function of FPR) for the proposed method ("Group Lasso", black solid lines), Lasso (dashed red lines) and Elastic-Net (green dotted lines) for $\rho \in \{0, 0.1, 0.2\}$. Top row: block-level evaluation; bottom row: SNP-level evaluation. The number of clusters for the proposed method is set to the true number of clusters.

For all positive values of correlation, the proposed method outperforms the other methods in the block-level evaluation. This is not surprising, as this method has been constructed to work in such situations. However, we believe that good performance at the block level is a very encouraging feature for GWAS.

For the SNP-level evaluation, the proposed method also outperforms Lasso and Elastic-Net as soon as $\rho \geq 0.2$. This is quite remarkable, because by construction the proposed method is bound to select either all SNPs of a block, or none, while there are at most 2 causal SNPs per block. Therefore, even if our proposed method is advantaged by the fact that it is given the true number of blocks, it is also a priori clearly disadvantaged by the SNP-level evaluation. We also observe that for specificities larger than 80% (that is, FPR $< 20\%$) there is virtually no difference between the Lasso and the Elastic-Net, suggesting that the correlation structure is too weak (for small but realistic values of $\rho$) for the grouping effect of the Elastic-Net to be effective. The grouping

effect of the proposed method is more effective because this method specifically looks for blocks of *adjacent SNPs*.

### 4.2 Varying the Number of Clusters

The setting considered in the preceding section may be seen as an Oracle setting for the proposed method, in the sense that the number of blocks was set to the true number of blocks. In practice however, the true number of blocks is unknown, and it currently has to be set by the user, so we have investigated the robustness of the proposed method to this parameter. Using the same simulation setting as above where the true number of clusters is 9, we have evaluated the proposed method when applied with 5 target clusters, corresponding to an "under-clustering" situation and with 13 target clusters, corresponding to an "over-clustering" situation. The results are shown in FIG. 2 for the under-clustering, and FIG. 3 for over-clustering. By construction, the performance of Lasso and Elastic-Net does not depend of the target number of clusters.

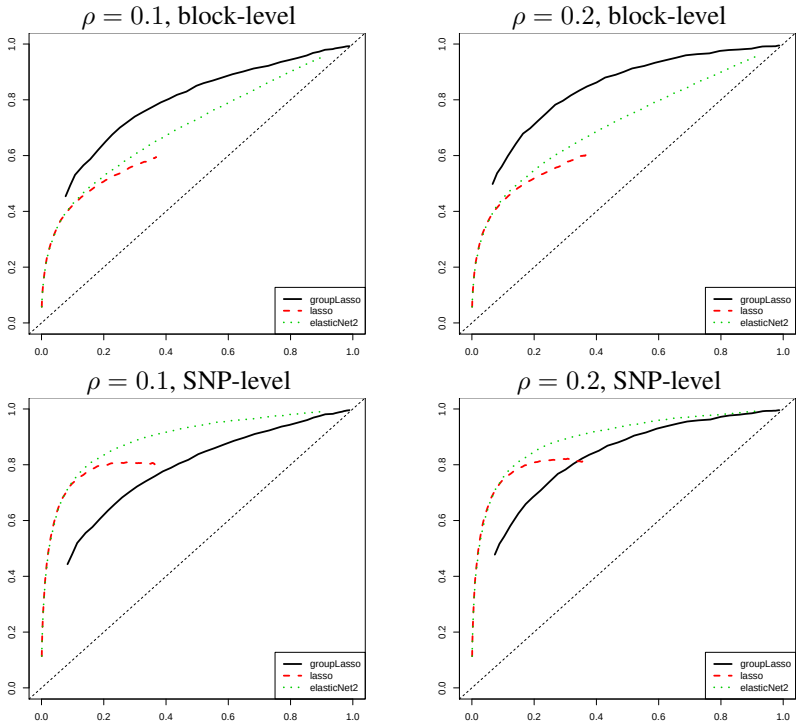



**Figure 2.** ROC curves (TPR in function of FPR) for the proposed method ("Group Lasso", black solid lines), Lasso (dashed red lines) and Elastic-Net (green dotted lines) for $\rho \in \{0.1, 0.2\}$. Top row: block-level evaluation; bottom row: SNP-level evaluation. The number of clusters for the proposed method is set to 5.

As in FIG. 1, the proposed two-step approach outperforms its competitors for block-level evaluation, both for under- and over-clustering. For SNP-level evaluation, the proposed method is outperformed by Lasso and Elastic-Net for under-clustering (FIG. 2). Indeed, with a target number of groups smaller than the actual one, the Group Lasso makes mistakes by canceling or activating too large groups. As for the over-clustering (FIG. 3) it is remarkable that the proposed method outperforms both Lasso and Elastic-Net: the Group Lasso does not suffer from over-clustering, as it can activate the "good" blocks among the ones clustered.

## 5 Conclusion and Perspective

In this paper, we have proposed a two-step approach that takes into account the biological information of the linkage disequilibrium between variables by firstly inferring LD blocks and then performing Group Lasso regression. State-of-the-art one-stage variable selection regression methods Lasso and Elastic-Net are outperformed by our proposed method for the purpose of identifying blocks containing causal SNPs, which we argue is a quite interesting feature in practice given the underdetermination of most GWAS. Interestingly, although the proposed method can only select groups of SNPs and not individual SNPs, it also achieves similar or better

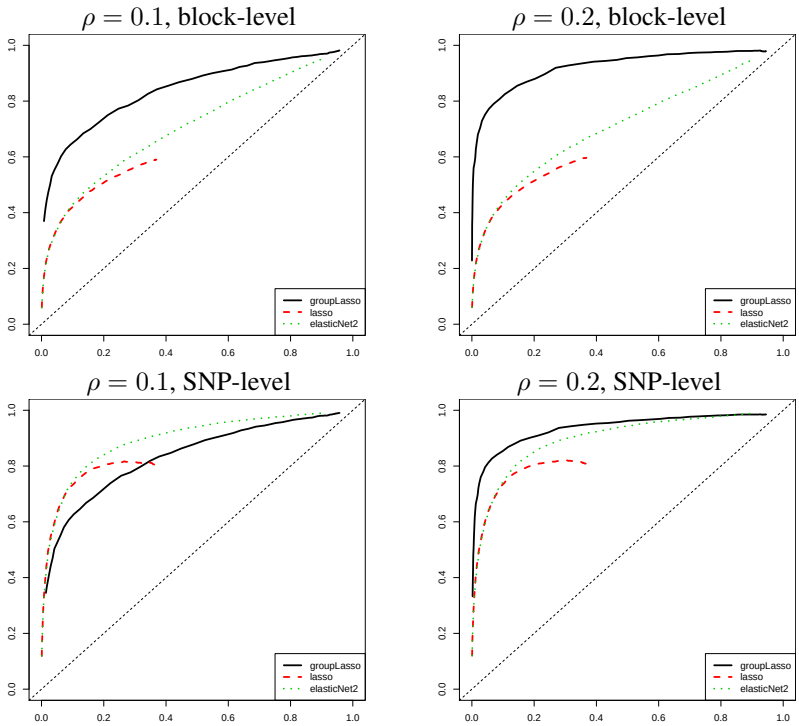



**Figure 3.** ROC curves (TPR in function of FPR) for the proposed method ("Group Lasso", black solid lines), Lasso (dashed red lines) and Elastic-Net (green dotted lines) for $\rho \in \{0.1, 0.2\}$. Top row: block-level evaluation; bottom row: SNP-level evaluation. The number of clusters for the proposed method is set to 13.

performance than its competitors in terms of selection of "causal SNPs". We believe that these results illustrate the relevance of the approach, and thereby the importance of tailored integration of biological knowledge in high-dimensional genomic studies such as GWAS.

A current limitation of the method is that it does not perform automatic model selection at the clustering stage, meaning that the user has to explicitly specify a target number of blocks. Our results in the case where the target number of blocks is misspecified (Section 4.2) are encouraging, as they show that the proposed method is fairly robust to situations where the target number of blocks is over-estimated. In order to improve this aspect of the method, several directions can be investigated. We will explore several model selection criteria that are adapted to the proposed constrained clustering. One alternative interesting possibility consists of replacing the Group Lasso by a *Hierarchical* Group Lasso. Another line of research considers replacing the current two-stage approach by a one-stage penalized regression method, by constructing a penalty that takes full advantage of the prior information that relevant groups of predictors can be expected to be adjacent along the genome.

## Acknowledgements

This work was partially funded by Institut National du Cancer (INCa) and Cancéropôle Ile-de-France.

The authors warmly thank Cyril Dalmasso and Julien Chiquet for very helpful discussions.

## References

[1] K. D. Bennett. Determination of the number of zones in a biostratigraphical sequence. *New Phytologist*, 132(1):155–170, 2006.

[2] P. Breheny and J. Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369, 2009.

[3] D. Clayton and H.-T. Leung. An R package for analysis of whole-genome association studies. *Human heredity*, 64(1):45–51, 2007.

[4] J. Liu, J. Huang, S. Ma, and K. Wang. Incorporating group correlations in genome-wide association studies using smoothed group lasso. *Biostatistics*, 2012.

[5] M. Slawski, W. Zu Castell, and G. Tutz. Feature selection guided by structural information. *The Annals of Applied Statistics*, 4(2):1056–1080, 2010.

[6] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[7] L. Wang, G. Chen, and H. Li. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494, 2007.

[8] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[9] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

[10] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2005.

[11] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

# Pathway mutation status predicts chemotherapy response in triple negative breast cancer

Magali MICHAUT[1,6], Esther H. LIPS[2,6], Lennart MULDER[2], Marlous HOOGSTRAAT[3], Marco J. KOUDIJS[3], René BERNARDS[1], Jelle WESSELING[4], Sjoerd RODENHUIS[5,7] and Lodewyk F. A. WESSELS[1,7]

[1] Department of Molecular Carcinogenesis, Netherlands Cancer Institute, Amsterdam, The Netherlands
{m.michaut, r.bernards, l.wessels}@nki.nl
[2] Department of Molecular Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands
{e.lips, l.mulder}@nki.nl
[3] Center for Personalized Cancer Treatment, Department of Medical Oncology UMC Utrecht Utrecht, The Netherlands
{M.Hoogstraat-2, M.J.Koudijs}@umcutrecht.nl
[4] Department of Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands
j.wesseling@nki.nl
[5] Department of Clinical Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands
s.rodenhuis@nki.nl
[6] Equally contributed
[7] Corresponding authors

**Abstract** *No targeted treatments exist for triple negative breast cancer, leaving chemotherapy as the only treatment option. Even though initial response to chemotherapy is often good, many patients relapse and develop resistance to chemotherapy. To identify biomarkers of chemotherapy resistance and putative directed treatment targets, we performed next generation sequencing of DNA from 31 pretreatment biopsies and matched normal blood. The patients received neoadjuvant chemotherapy and were divided in responders and non-responders, depending on whether or not a pathological complete remission was achieved. Focusing on somatic mutations, we found that only a small number of genes were mutated in several samples and none of them were predictive of response. However, a pathway analysis showed that mutations in phosphatidylinositol signaling (PI3K) were significantly more frequent in the non-responders and predicted to be damaging. After validation, treatment regimens that combine chemotherapy with blockage of this pathway should be investigated for these tumors.*

**Keywords** Breast cancer, chemotherapy resistance, mutations, PI3K pathway

## 1 Background

Primary triple negative breast cancers (TNBC) are defined by a lack of estrogen receptor (ER), progesterone receptor (PR) and Her2 gene amplification. They represent 16% of all breast cancers [1]. No targeted treatments exist for TNBCs, leaving conventional chemotherapy as the only treatment option. Even though initial response to chemotherapy is often good, many patients relapse and develop resistance to chemotherapy. Consequently it is essential to better understand which patients respond well or not to chemotherapy, and why they do so.

Being able to predict chemotherapy response is important to avoid negative effects of chemotherapy for the patients who would not benefit from this treatment. In addition, it is required to find alternative treatment options for these patients resistant to chemotherapy. In this project, we investigate the link between molecular alterations and response to chemotherapy in TNBCs in order to identify biomarkers of chemotherapy resistance and putative directed treatment targets. We focus here on genetic variants and perform high-throughput sequencing.

## 2 Results

We performed next generation sequencing for 2000 genes involved in various pathways known to contain genes mutated in cancer, using the Mini Cancer Genome library [2]. DNA from 31 pretreatment biopsies and

matched normal blood was sequenced. Biopsies were derived from patients scheduled to receive neoadjuvant chemotherapy (doxorubicin and cyclophosphamide). Tumors were divided in responders (n=15) and non-responders (n=16), depending on whether or not a pathological complete remission was achieved on breast and lymph nodes.

Focusing on somatic mutations, we first analysed each gene separately. Only a small number of genes (n=9) were mutated in several samples and none of them were predictive of response. Thus we decided to combine the low-frequency mutations based on the pathway they target. Based on a pathway analysis, we found that mutations in the *phosphatidylinositol signaling* (PI3K) pathway were significantly more frequent in the non-responders with mutations present in 11/16 non-responders and 1/15 responders. Similar level of enrichment was found in the *inositol phosphate metabolism pathway*. Interestingly most mutations were predicted as damaging and/or present in the COSMIC database, supporting the idea that these mutations are functional and important for the tumor.

A possible bias would be that the responders and non-responders have different mutation rates overall. In that case, the pathways well represented in the set of genes we sequenced could be found significant as false positive. However we confirmed that responders and non-responders have similar mutation rates. On average, samples had 10 mutations, corresponding to a mutation rate of 1.8 somatic mutations per Mb, which is similar to previous rates found [3]. This confirms that non-responders tend to have more mutations in the PI3K pathway.

The next step is to find out which treatment could target specifically these tumours mutated in PI3K pathways and not responding well to chemotherapy. To study this, we took benefit of the large-scale drug sensitivity panel provided by the Sanger Institute [4]. In this panel, a total of 716 cell lines (including 43 breast cancer lines) were tested for sensitivity against 138 drugs (targeted and cytotoxic). For each drug, we compared the sensitivity of cell lines mutated in PI3K pathway and other cell lines. The most significant associations we found are for PI3K inhibitors: cell lines mutated in this pathway are more sensitive. This is not surprising but supports the idea that a combination of chemotherapy and PI3K inhibitor may be beneficial for patients not responding to chemotherapy only.

## 3 Conclusion

We found that mutations in genes of the PI3K pathway occur frequently in triple negative breast cancers that do not achieve a pathological complete remission on neo-adjuvant chemotherapy with doxorubicin and cyclophosphamide. Moreover, cell lines mutated in PI3K pathway are more sensitive to PI3K inhibitors than non-mutated cell lines. Thus treatment regimens that combine chemotherapy with blockage of this pathway should be investigated for these tumors. If the result is further validated, this could have a major clinical impact.

One of the challenges of high-throughput sequencing is that numerous genes have mutations in only few samples. Methods based on highly frequently mutated genes are not enough to understand the effect of all these low-frequency mutations and the interplay between them. This work shows the benefit of pathway analysis to group together mutations which target the same pathway and get an integrated information.

In conclusion, this work has two major implications: i) the biological findings which may lead to an important clinical outcome ii) the methodology used which enables to benefit from high-throughput sequencing experiments even for a small number of samples.

## References

[1] Blows, Fiona M and Driver, Kristy E and Schmidt, Marjanka K and Broeks, Annegien and van Leeuwen, Flora E and Wesseling, Jelle and Cheang, Maggie C and Gelmon, Karen and Nielsen, Torsten O and Blomqvist, Carl and Heikkilä, Päivi and Heikkinen, Tuomas and Nevanlinna, Heli and Akslen, Lars A and Bégin, Louis R and Foulkes, William D and Couch, Fergus J and Wang, Xianshu and Cafourek, Vicky and Olson, Janet E and Baglietto, Laura and Giles, Graham G and Severi, Gianluca and McLean, Catriona A and Southey, Melissa C and Rakha, Emad and Green, Andrew R and Ellis, Ian O and Sherman, Mark E and Lissowska, Jolanta and Anderson, William

F and Cox, Angela and Cross, Simon S and Reed, Malcolm W R and Provenzano, Elena and Dawson, Sarah-Jane and Dunning, Alison M and Humphreys, Manjeet and Easton, Douglas F and García-Closas, Montserrat and Caldas, Carlos and Pharoah, Paul D and Huntsman, David, Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS medicine*, 7:e1000279, 2010.

[2] Vermaat, Joost S and Nijman, Isaäc J and Koudijs, Marco J and Gerritse, Frank L and Scherer, Stefan J and Mokry, Michal and Roessingh, Wijnand M and Lansu, Nico and de Bruijn, Ewart and van Hillegersberg, Richard and van Diest, Paul J and Cuppen, Edwin and Voest, Emile E, Primary colorectal cancers and their subsequent hepatic metastases are genetically different: implications for selection of patients for targeted treatment. *Clinical Cancer Res.*, 18:688-699, 2012.

[3] Banerji, Shantanu and Cibulskis, Kristian and Rangel-Escareño, Claudia and Brown, Kristin K and Carter, Scott L and Frederick, Abbie M and Lawrence, Michael S and Sivachenko, Andrey Y and Sougnez, Carrie and Zou, Lihua and Cortes, Maria L and Fernandez-Lopez, Juan C and Peng, Shouyong and Ardlie, Kristin G and Auclair, Daniel and Bautista-Piña, Veronica and Duke, Fujiko and Francis, Joshua and Jung, Joonil and Maffuz-Aziz, Antonio and Onofrio, Robert C and Parkin, Melissa and Pho, Nam H and Quintanar-Jurado, Valeria and Ramos, Alex H and Rebollar-Vega, Rosa and Rodriguez-Cuevas, Sergio and Romero-Cordoba, Sandra L and Schumacher, Steven E and Stransky, Nicolas and Thompson, Kristin M and Uribe-Figueroa, Laura and Baselga, Jose and Beroukhim, Rameen and Polyak, Kornelia and Sgroi, Dennis C and Richardson, Andrea L and Jimenez-Sanchez, Gerardo and Lander, Eric S and Gabriel, Stacey B and Garraway, Levi A and Golub, Todd R and Melendez-Zajgla, Jorge and Toker, Alex and Getz, Gad and Hidalgo-Miranda, Alfredo and Meyerson, Matthew, Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 486:405-409, 2012.

[4] Garnett, Mathew J and Edelman, Elena J and Heidorn, Sonja J and Greenman, Chris D and Dastur, Anahita and Lau, King Wai and Greninger, Patricia and Thompson, I Richard and Luo, Xi and Soares, Jorge and Liu, Qingsong and Iorio, Francesco and Surdez, Didier and Chen, Li and Milano, Randy J and Bignell, Graham R and Tam, Ah T and Davies, Helen and Stevenson, Jesse A and Barthorpe, Syd and Lutz, Stephen R and Kogera, Fiona and Lawrence, Karl and McLaren-Douglas, Anne and Mitropoulos, Xeni and Mironenko, Tatiana and Thi, Helen and Richardson, Laura and Zhou, Wenjun and Jewitt, Frances and Zhang, Tinghu and O'Brien, Patrick and Boisvert, Jessica L and Price, Stacey and Hur, Wooyoung and Yang, Wanjuan and Deng, Xianming and Butler, Adam and Choi, Hwan Geun and Chang, Jae Won and Baselga, Jose and Stamenkovic, Ivan and Engelman, Jeffrey A and Sharma, Sreenath V and Delattre, Olivier and Saez-Rodriguez, Julio and Gray, Nathanael S and Settleman, Jeffrey and Futreal, P Andrew and Haber, Daniel A and Stratton, Michael R and Ramaswamy, Sridhar and McDermott, Ultan and Benes, Cyril H, Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483:570-575, 2012.

# Gene containing Variant Annotation for Prioritization
# a tool guiding clinicians toward candidate variations of interest

Nadia Bessoltane[1], Virgine Bernard[1], Olivier Delattre[2,3]

[1] Institut Curie, plateforme Next Generation Sequencing, 26 rue d'Ulm, 75248 Paris cedex 05, France
{nadia.bessoltane-bentahar, virginie.bernard}@curie.fr

[2] Institut Curie, unité de génétique somatique, 26 rue d'Ulm, 75248 Paris cedex 05, France

[3] Institut Curie, INSERM U830, laboratoire de génétique et biologie des cancers, 26 rue d'Ulm, 75248 Paris cedex 05, France
olivier.delattre@curie.fr

**Abstract** *Current software allow to predict variations within protein encoding genes that are likely to contribute to a disease phenotype, by annotating them. Taking advantage of these functional and structural annotations, current analysis mainly focuses on novel nonsense and missense variants. With the emergence of full-genome analysis, the variations outside of exons are accessible and are going to be reported. They play a role in disorders and should not be filtered. The list of variant to validate will grow-up requiring to prioritize the ones the more likely to be of interest. Gene containing variant Annotation for Prioritization, GeVAP, is a new tool allowing further annotations leading scientists to highlight the gene containing variants that are likely to play a role in the disease analysed.*

**Keywords** *Next Generation sequencing, Causal variant prediction, Variant prioritization, Gene annotation.*

## 1 Introduction

The convergence of high-throughput technologies for sequencing individual exomes and full-genomes and rapid advances in genome annotation are driving a neo-revolution in human genetics. The identification of causal variations is accessible on a large scale, quickly, and is affordable. By mapping the reads obtained by Next Generation Sequencing (NGS) to the human genome reference and by searching for variations relative to the reference, a list of small nucleotide variations, insertions, deletions as well as structural rearrangements is obtained [1,2,3]. Aligners assess the quality of predictions and consider whether they are likely to be real or false positive. Variant structural annotation with a reference genome allows to focus on variations within protein-coding genes in first intention [4]. Then, software and annotation pipelines allow filtering the changes that are most likely to be deleterious and contribute to diseases. First, it is informative to highlight new variations, i.e. unknown in polymorphism databases and former sequencing analysis [5,6]. Then variations that are known to be somatic and listed in the COSMIC database (Catalogue Of Somatic Mutations In Cancer) can be reported [7]. Finally, existing software score the severity of single substitutions that lead to a missense or a nonsense variation by assessing single amino acid substitution impact on protein structure [8,9,10]. All these steps reduce the number of false positive variations and then the list of variation of interest. At the end of NGS analysis, once all above filters applied, the list of candidate variations predicted is submitted for validation. Most approaches consist in first intention in the validation on all novel non-sense and missense variation within protein-encoding exons.

Nevertheless, protein-encoding exons are only 2% of the genome and the remaining 98% of the genome control the developmental and physiological profile of gene activity - when and where a gene will be active. These regions are known to be linked to disease phenotype [11,12,13,14,15]. Current list of candidate variant is led to multiply with whole genome analysis will be done routinely, likely tomorrow. Indeed, functional contributions of cis-regulatory sequence variations to human genetic disease are numerous. The prediction of these causal regulatory variations already started [16,17]. More generally, it is imperative no more to focus on protein-encoding exons but to look at all variations, the ones involved in splicing being already accessible during exome analysis. The list of variations predicted of interest is going to grow-up in the next few months.

Variant validations are time-consuming and expensive. To highlight on gene containing variants that make sense relative to the disease phenotype studied would help to prioritize causal variations. To gain

biological understanding from Next Generation Sequencing (NGS) analysis is required and accessible. Software allow analyzing the functional annotations of gene lists. Deeply used, DAVID performs gene-enrichment and functional annotation analysis and gene functional classification [18,19]. Nevertheless, DAVID has not been designed for processing high-throughput technologies data. NGS analysis needs dedicaced tools. To our knowledge, CaGe is the only web-accessible tool aiming gene annotation from NGS-based analysis. It is dedicaced for Cancer Genomics [20]. Not only for cancer, gene annotation is required to help scientists rank predicted variations with respect to their potential to be causal for the disease. Easier than web-based tools, an automatic functional annotation of genes usable in command line into an NGS pipeline is accessible. This is the purpose of the gene containing variant Annotation for Prioritization: GeVAP.

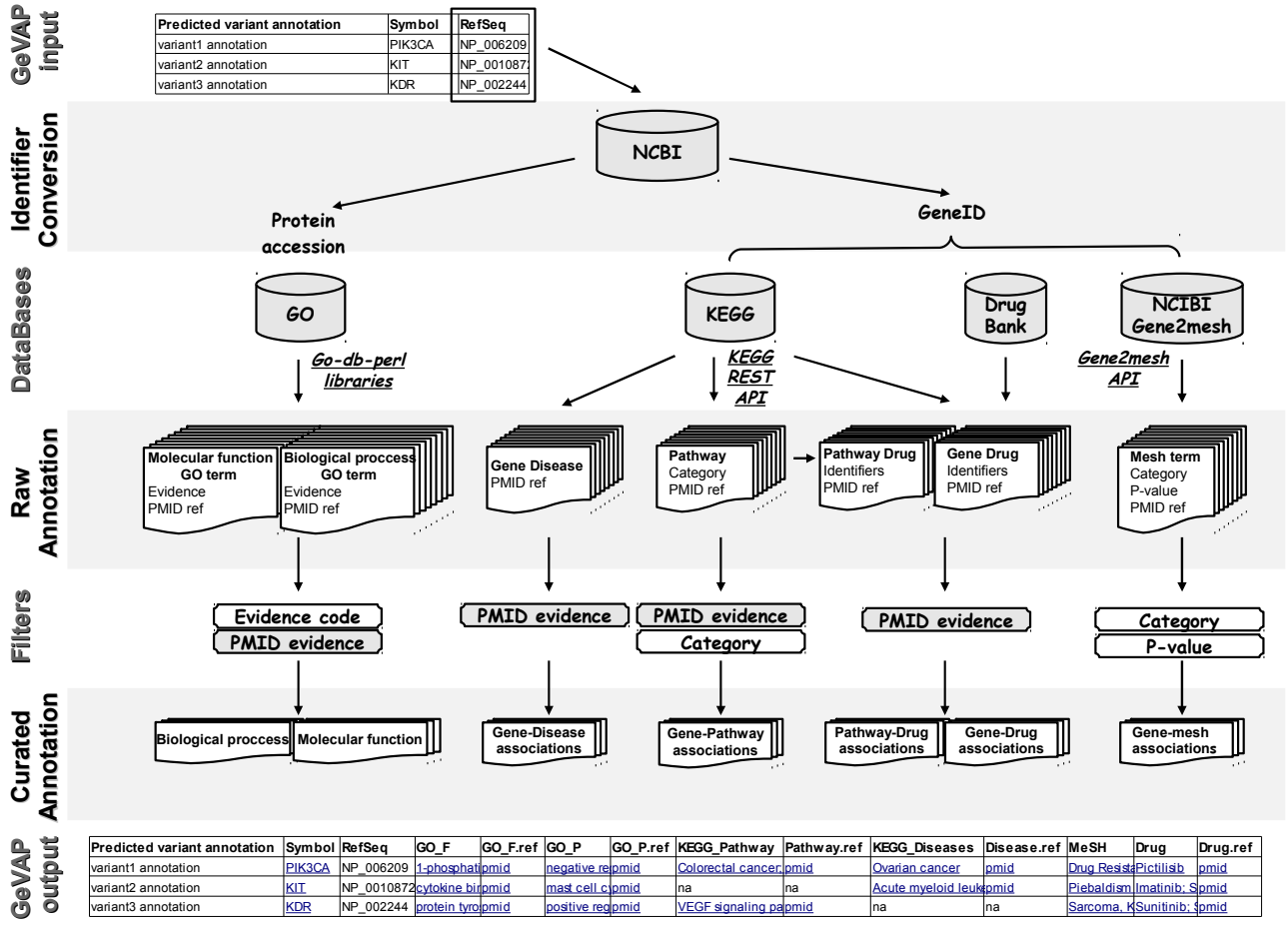



**Figure 1.** Gene containing Variant Annotation for Prioritization workflow

Minimum input information required by GeVAP is RefSeq unique protein isoform identifiers. As filters, GeVAP used Gene Ontology (GO) evidence codes to focus only on experimental annotations. It analyses literature to report accurate PubMed Identifier (PMID) evidences. By default, only disease categories are reported. Gene2MeSH p-value provided by default on the web site filters accurate gene-MeSH annotations. Output format reports curated gene annotations in addition to variant annotations, and provides links to database web pages and pubmed for evidences.

## 2 Material and methods

Gene containing variant Annotation for Prioritization (GeVAP) provides annotation of genes as detailed below and as summarized on the tool workflow Figure 1.

### 2.1 Literature evidences for gene annotation

In order to focus on curated gene annotations we developed an approach filtering genes annotations supported by publication(s). Through NCBI tools, GeVAP gets all pubmed identifiers (PMID) linked to a gene [21]. This step is done once per gene leading to a gene PMID list. Then each article reported by used-databases as evidence of gene annotation is checked. Comparison of the annotation evidence PMID list and the gene PMID list is done. If shared PMID are observed, the gene annotation is reported and PMID are reported as evidence. Otherwise, the annotation is not reported, PMID evidence missing (Figure 2).

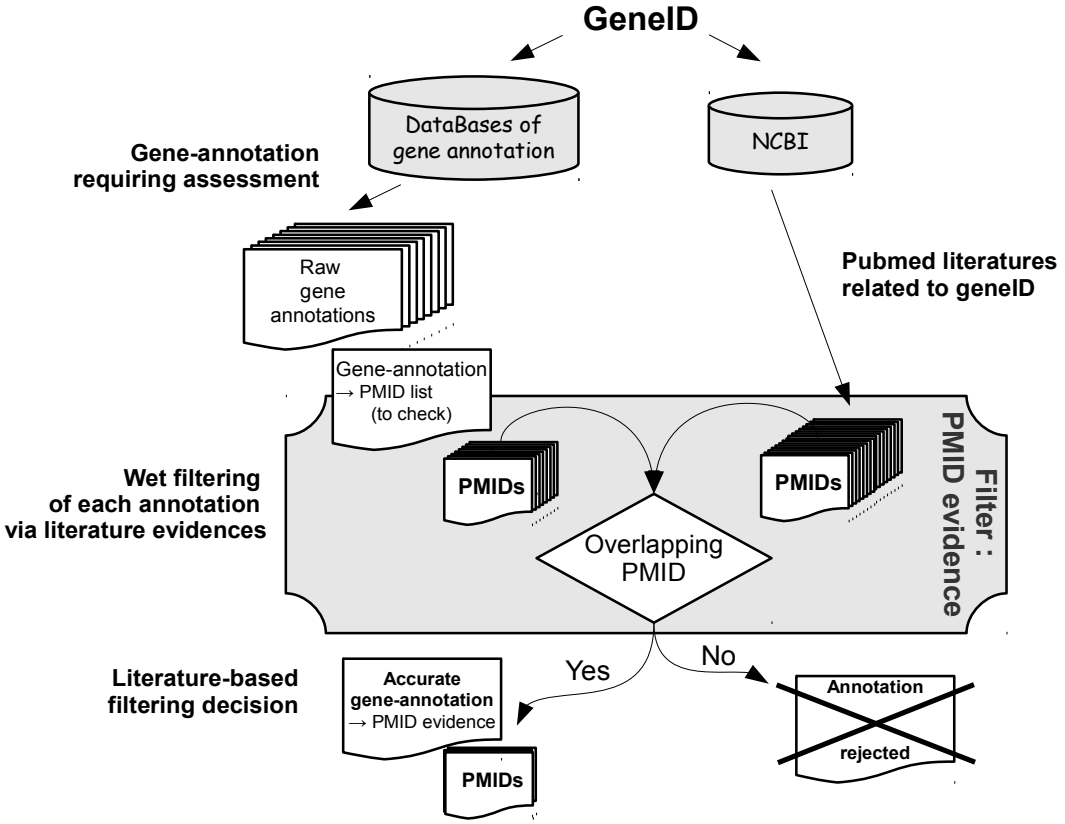



**Figure 2.** Literature-based filtering (PMID evidence) workflow
GeVAP reports only accurate annotation supported by PMID(s).

### 2.2 Gene identifier converter

GeVAP support as input RefSeq protein identifiers. In order to request the databases, our tool uses this unique gene product identifiers to retrieve gene identifiers (entrez gene) and protein accession numbers from NCBI [22] (Figure 1).

### 2.3 Gene molecular function and biological process annotations

The Gene Ontology (GO) website provides cellular component, genes molecular function and biological process annotation [23,24]. Through GeVAP gene ID converter tool, the required protein accession number is obtained (See "Gene identifier converter") and used for database requesting. For each protein ID, biological process and molecular function annotations are extracted from GO database. In addition, in order to allow annotation quality filtering, evidence codes and their references are requested. They distinguish annotations provided by experimental analysis, computational analysis, author statements, and other none accurate annotation. By default, we focus on curated molecular function and biological process annotation, i.e. the ones having experimental evidences. Then the literature-based filtering is irrelevant. It is implemented but not used by default due to evidence code filtering. It is useful otherwise. As output, evidence codes and

pubmed identifiers are reported as annotation evidences.

## 2.4 Gene pathway annotations

The Kyoto Encyclopedia of Genes and Genomes (KEGG) reports in a database the knowledge on gene molecular interaction and reaction networks [25,26]. KEGG provides a list of publication as evidence of genes-pathway link. In order to distinguish between the manual curated and automatic deduction of gene-pathway association, GeVAP filters literature evidences as described above (See "Literature evidences for gene annotation"). KEGG divides gene-pathways in six categories (i) Metabolism, (ii) Genetic information processing, (iii) Environmental information processing, (iv) Cellular processes, (v) Organismal systems, and (vi) Human Diseases. In addition to pathway filtering via PMID evidences, by default, GeVAP focus on pathways involving human diseases, the main focus of human variant calling projects. This filter can be updated depending on projects scientists are working on. All the categories being divided in smaller subsets, it allows flexibility in filtering.

## 2.5 Gene-related disease annotations

The KEGG database reports existing genes-related diseases [27]. As annotation and evidences, KEGG provides a description of disease and a list of PMID. As for pathways, GeVAP apply literature-based filtering pipeline to report accurate annotations.

## 2.6 Genes Medical Subject Heading terms annotations

Medical Subject Heading terms (MeSH terms) are the National Library of Medicine's controlled vocabulary thesaurus used for indexing articles for PubMed [28]. The Gene2MeSH allows MeSH searching using as query a gene symbol or gene ID [29]. A significant association of a gene and a MeSH-term is assessed via a Fisher's Exact Test calculating a p-value. The p-value takes into account the occurrence number of each MeSH term in the literature avoiding false positive association for recurrent terms. As results, Gene2Mesh provides for each gene/MeSH term a p-value and a list of PMID. GeVAP uses as p-value threshold default value proposed by Gene2MeSH.

MeSH terms are divided into sixteen categories, including "Diseases" and "Chemicals and Drugs" [30]. Categories are sub-divided into heading. MeSH terms being organized in acyclic graph, one term can belong to more than one category or heading. By default, GeVAP select MeSH terms of Neoplasms heading from Disease category [31]. This filter can be suited to the project.

## 2.7 Drug inhibiting-gene and inhibiting-pathway annotations

The KEGG database links genes and pathways to their drug inhibitors [32]. Nevertheless, no public evidence based on literature is accessible. Experimental evidence filtering as done for GO cannot be applied on these annotations. Then as a complement, information liking directly genes and drugs is provided by DrugBank [33,34]. Experimental evidence filtering using PMID is done (See "Literature evidences for gene annotation" session above) and combined between KEGG and DrugBank information leading to curated pathway-drug and gene-drug associations.

## 2.8 Gene containing variant Annotation for Prioritization implementation

GeVAP is implemented in PERL [35]. Access to databases and their tools, used for annotation, is done remotely through specific Application Programming Interfaces (API) for each database. GeVAP uses the API provided by go-db-perl library to request the Gene Ontology database. It uses the KEGG REST-based API service for customization querying in KEGG databases. The output of all operations is in a text format. The 'LWP' modules (standard Perl modules) are used to access to web requests and the outputs are parsed with Perl classical scripts. The National Center for Integrative Biomedical Informatics (NCIBI) provides API access to Gene2MeSH tool. To access to NCBI Gene and PubMed databases GeVAP use 'Entrez Programming Utilities' (E-utilities: version 2.0). This E-utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve the requested data [36]. The output of their tow API (Gene2MeSH programmatic interface and E-utilitie) is in XLM format. To parse and retrieve data GeVAP use Perl 'XML::LibXML' [37].

# 3 Results

## 3.1 Gene containing variant Annotation for Prioritization filtering step : a mandatory step leading to accurate annotations

GeVAP aims helping users to quickly prioritize candidate genes from NGS-bases variant calling analysis. It uses gene annotation provided in public distant databases to retrieve updates gene annotations. GeVAP filters curated annotations based on wet evidences and provided PMID list as evidences. Indeed, by default, databases provide many gene-annotations not all accurate. Users have to filter. GO database allows to filter based on experimental evidences. Raw molecular function and biological process annotations including any evidences lead respectively to 15 026 and 14 545 genes annotated, some genes having up to 135 annotations. Most of these annotations are predicted and then irrelevant for GeVAP purpose. Once GeVAP PMID evidence filters applied, 7 087 and 5 567 genes are annotated.

KEGG pathway maps are manually curated on the basis of the experimentally obtained knowledge about some specific organisms. Then, it is automatically reconstructed across other organisms by the KEGG orthology system. Filters are required to focus on accurate gene-pathway annotations. KEGG provides as evidences for gene annotations a list of publications, some articles do not concerning the gene. KEGG evidences need to be check in order to focus on experimental evidences (see “Literature evidences for gene annotation” in Material and Methods for details). Raw KEGG pathway and disease annotations without literature-based filtering lead respectively to 6 430 and 2 240 genes annotated. Once GeVAP filters applied, 617 and 986 genes are annotated.

GeVAP curated annotation provided are (i) gene molecular function and biological process, (ii) gene pathway, (iii) gene diseases, (iv) gene and pathway drug inhibitor and (v) Medical Subject Heading. GeVAP experimental and literature-based filtering is important for accurate annotation reporting (Figure 2). Aiming to help variant prioritization, only accurate information should to be reported. That is the purpose of default values used. Many genes will not be annotated, but it is likely better than to provide users many annotations, some without evidences.

## 3.2 Annotation performed on gene isoforms

Common variant structural annotation software provide the gene symbol where a predicted variant is located [9,8]. In addition, most of them report the accession numbers of gene products affected by variations and variant annotations following the Human Genome Variation Society nomenclature of variations (HGVS). Indeed, one variant may be related to several isoforms of the same gene and isoforms may have different functions. In order to exhaustively annotate genes containing a variant, all isoforms are analysed. Refseq identifiers unique to each isofome and gene symbol are used to further request databases for annotation. GeVAP gene identifier converter functions allow to provide all databases the input they need for further analysis. In addition, it allows flexibility for users who may provide as input file many variant annotated file containing any gene identifier.

GeVAP requires as input variant file annotated and containing one column for gene identification. RefSeq of protein or mRNA are recognized. SIFT or polyphen-2 output as any tab-separated file are recognized and parsed by GeVAP to extract the list of gene to be annotated. It requests gene annotating databases, and extract curated annotation. GeVAP reports as output file the initial input file with additional columns providing gene annotations and links to annotations and their evidences. GeVAP allow straightforward data processing (Figure 1).

## 3.3 User friendly report of variant and gene annotations

Gene annotation of interest to better highlight if it make sens hypothesis a gene might be involved in a disorder are (i) gene-function, (ii) gene-pathways and (iii) gene-MeSH terms, i.e. controlled vocabulary thesaurus used for indexing literature. For this purpose, GeVAP uses three databases. The Gene Ontology (GO) provides structured, controlled vocabularies and classifications for several domains of molecular and cellular biology. GeVAP ranks annotations by evidence codes that describe the work or analysis upon which the gene-function association is based (See Material and Methods). Experimental evidence codes related with pubmed article references that support this work are the ones GeVAP reports by default. To report gene-pathways, GeVAP request KEGG database, a collection of manually drawn pathway maps representing the

knowledge on the molecular interaction and reaction networks. Gene-pathway evidences are based on filtered literature (Figure 2). Finally, GeVAP requests gene2Mesh to report MeSH-terms significantly overrepresented in a gene literature. In addition, literature leading to the link gene and MeSH term is reported.

Moreover, once a variant has been filters as candidate involved in a disorder, knowing drugs that can be used to enhance well-being of disabled patients is powerful information clinicians are interested by. Indeed, some genes and some pathways are inhibited by chemical substances. A drug inhibiting a gene involved in the same pathway than a mutated gene may be effective though not linked to the gene containing variant directly. In addition of providing gene-pathway, KEGG reports drugs inhibiting pathways [27]. A drug inhibiting a gene involved in the same pathway than a mutated gene might be effective though not linked to the gene of interest itself. Complementary to KEGG drug-pathway annotation, the DrugBank database provides detailed information about drugs and their targets and the literature about them. GeVAP reports gene-drugs and pathways-drugs once literature filtering done.

Aiming to provide user friendly reports of variant and gene annotation for biologists, GeVAP generate a tabulated file Excel or Open Office can open. It reports input information and accurate (i) gene molecular function and biological process, (ii) gene pathways, (iii) gene MeSH-terms, and (iv) gene and pathway inhibiting-drugs all with PMID evidence. Annotations are reported in two columns: gene-annotations with a link to annotation description web site, and PMID with a link to pubmed web site listing all literature evidences. Only gene-MeSH annotations are reported in one column dedicated with the gene-Mesh terms and a link to pubmed web site highlighting PMID evidences. For each gene, GeVAP added a link to GeneCards providing detailed information about gene containing variant [34]. The link to online databases provides all details users may need to dig once an annotation is of interest (Figure 3). Based on the assumption that GeVAP report is dedicaced to clinicians and scientists having a good knowledge of the disease affection patient analysed, the tool will enable them to benefit from gene annotation leading to variant prioritization. In addition GeVAP is a command line tool. It can be integrated to a pipeline making it easy to use for computer scientists.

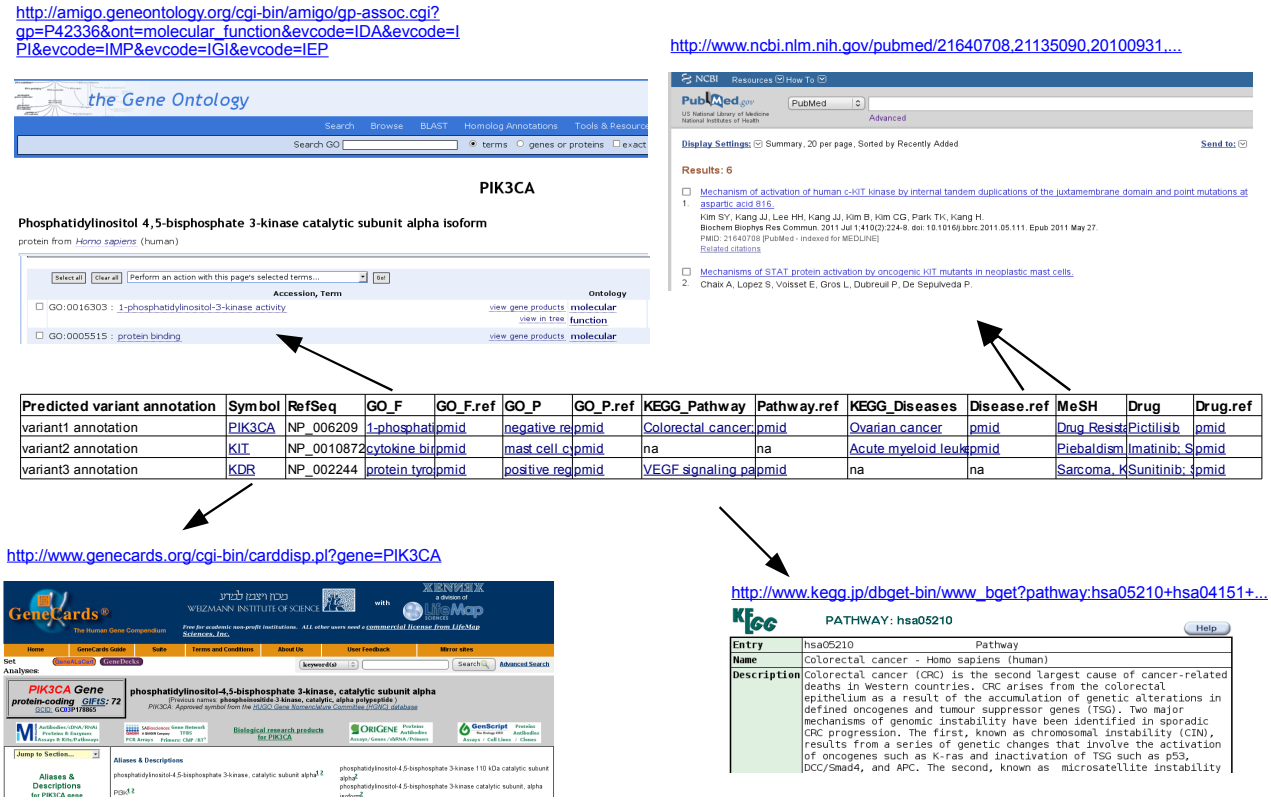



**Figure 3.** Output provided by Gene containing variant Annotation for Prioritization (GeVAP).

Output file reports all input file information completed by further gene accurate annotations with PMID evidences. All annotation is reported via a link to further details on web pages and via a link to pubmed literature evidences.

## 4 Conclusions

High-throughput sequencing in the past few years allowed detecting variations within protein encoding genes contributing to multiple disease phenotypes. Technologies improvements as well as in silico software development are going to report increasingly candidate variants. Though sensitivity as specificity are improved, sanger is required for validation. This validation is time-consuming and expensive. A tool prioritizing variants to validate that are likely to be involved in disorder is required. To our knowledge, gene containing variants are only annotated based on structural information following HGVS nomenclature [4], polymorphism and frequency analysis [5,6] and damaging impact [7,8,9,10]. Further annotation of gene is required and accessible. GeVAP is an ongoing project providing gene function, gene pathway and gene-targeted drugs annotation. It uses API allowing fast requests on updated data and generate output file providing all accurate annotations. GeVAP filters them based on experimental work. It provides a link to online databases and reports all annotation evidences via a link to online pubmed article reporting gene-annotation (Figure 3).

Thanks to the output file providing all links to databases and evidences, clinicians can access further information for annotations on interest.

A good knowledge of the disease under study is necessary to enable the best use annotation produced by GeVAP. The tool is dedicaced for clinicians. By using knowledge the patient-related disease, on first intention, gene functional annotation provided by GO, KEGG and gene2MeSH may highlight gene containing variants that make sense relative to the disease phenotype studied and then help to prioritize causal variations. Once validations done, drug information may provide information leading to clue how to improve patient well being via appropriate drugs.

GeVAP is an ongoing project likely to improve via further appropriate annotations depending on clinicians needs. It aims to improve variant prioritizing. GeVAP should allow user to provide patient disorder information and based on annotation and these disease information rank predicted variations.

## Abbreviations

API: Application Programming Interface; COSMIC: Catalogue of Somatic Mutations in Cancer; GeVAP: Gene containing Variant Annotation for Prioritization; GO: Gene Ontology; HGVS: Human Genome Variation Society; KEGG: Kyoto Encyclopedia of Genes and Genomes; MeSH: Medical Subject Headings; NGS: Next-generation sequencing; PMID: PubMed identifier.

## Author's contribution.

All authors contributed to the design of the method and the analysis and interpretation of the data. NB implemented and carried out the study. All authors read and approved the final manuscript.

## Acknowledgements

This work was supported by the The French National Research Agency under the program “Investments for the future” with the reference ANR-10-03-EQPX.

## References

[1] Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60.

[2] 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. 2010 Sep;20(9):1297-303.

[3] Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, Nicolas A, Delattre O, Barillot E. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics. 2010 Aug 1;26(15):1895-6.

[4] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep;38(16):e164.

[5] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of

genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11.

[6] Overbeek R et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005 Oct 7;33(17):5691-702.

[7] Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008 Apr;Chapter 10:Unit 10.11.

[8] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;Chapter 7:Unit7.20.

[9] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81.

[10] Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011 Aug;32(8):894-9.

[11] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248-9.

[12] VanderMeer JE, Ahituv N. cis-regulatory mutations are a genetic cause of human limb malformations. Dev Dyn. 2011 May;240(5):920-30.

[13] Epstein DJ. Cis-regulatory mutations in human disease. Brief Funct Genomic Proteomic. 2009 Jul;8(4):310-6.

[14] Garone C, Pippucci T, Cordelli DM, Zuntini R, Castegnaro G, Marconi C, Graziano C, Marchiani V, Verrotti A, Seri M, Franzoni E. FA2H-related disorders: a novel c.270+3A>T splice-site mutation leads to a complex neurodegenerative phenotype. Dev Med Child Neurol. 2011 Oct;53(10):958-61.

[15] Murphy SM, Polke J, Manji H, Blake J, Reiniger L, Sweeney M, Houlden H, Brandner S, Reilly MM. A novel mutation in the nerve-specific 5'UTR of the GJB1 gene causes X-linked Charcot-Marie-Tooth disease. J Peripher Nerv Syst. 2011 Mar;16(1):65-70.

[16] Worsley-Hunt R, Bernard V, Wasserman WW. Identification of cis-regulatory sequence variations in individual genome sequences. Genome Med. 2011 Oct 10;3(10):65.

[17] Shah SP et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012 Apr 4;486(7403):395-9.

[18] Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009 Jan;37(1):1-13.

[19] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.

[20] Park YK, Kang TW, Baek SJ, Kim KI, Kim SY, Lee D, Kim YS. CaGe: A Web-Based Cancer Gene Annotation System for Cancer Genomics. Genomics Inform. 2012 Mar;10(1):33-9.

[21] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2013 January; 41(Database issue): D8–D20.

[22] Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2011 Jan;39(Database issue):D52-7.

[23] Gene Ontology Consortium. Gene Ontology annotations and resources.Nucleic Acids Res. 2013 Jan;41(Database issue):D530-5.

[24] Hill DP, Smith B, McAndrews-Hill MS, Blake JA. Gene Ontology annotations: what they mean and where they come from. BMC Bioinformatics. 2008 Apr 29;9 Suppl 5:S2.

[25] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 1999 Jan 1;27(1):29-34.

[26] Tanabe M, Kanehisa M. Using the KEGG database resource. Curr Protoc Bioinformatics. 2012 Jun;Chapter 1:Unit1.12.

[27] Kanehisa M. Molecular network analysis of diseases and drugs in KEGG. Methods Mol Biol. 2013;939:263-75

[28] Chapter 11 Relationships in Medical Subject Headings. http://www.nlm.nih.gov/mesh/meshrels.html

[29] Ade, AS; Wright, ZC; States, DJ; Gene2MeSH [Internet]. Ann Arbor (MI): National Center for Integrative Biomedical Informatics. 2007 Mar. Available from http://gene2mesh.ncibi.org

[30] http://www.ncbi.nlm.nih.gov/mesh/1000048

[31] http://www.nlm.nih.gov/cgi/mesh/2013/MB_cgi?mode=dcms&term=Neoplasms&field=entry#TreeC04

[32] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008 Jan;36(Database issue):D901-6.

[33] Wishart DS. DrugBank and its relevance to pharmacogenomics. Pharmacogenomics. 2008 Aug;9(8):1155-62.

[34] Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota Madi A, Olender T, Golan Y, Stelzer G, Harel A, Lancet D. GeneCards Version 3: the human gene integrator. Database (Oxford). 2010 Aug 5;2010:baq020

[35] http://www.perl.org/

[36] http://www.ncbi.nlm.nih.gov/books/NBK25501/

[37] http://www.cpan.org/

# Session 4 : Biologie de synthèse et structure des protéines

## Conférence invitée

JEAN-LOUP FAULON

Institute of Systems & Synthetic Biology,
CNRS, Université d'Evry Val d'Essonne,
Genopole campus 1, 5 rue Henri Desbruères,
91030 Evry Cedex, France
Jean-Loup.Faulon@issb.genopole.fr, www.issb.genopole.fr/~faulon

# A Rational Metabolic Engineering Pipeline: from CAD to product identification

Synthetic biology addresses problems such as engineering metabolic pathways into chassis organisms to produce desired chemicals. With the goal of rationalizing the engineering process, I will present a suite of computational tools coupled with experimental validations. Metabolic reactions linking target chemicals to endogenous chassis metabolites are first searched by using retrosynthesis, a concept originally developed for synthetic chemistry. Genes encoding enzymes catalyzing metabolic reactions are then searched genome-wide using a machine learning technique developed in our group to predict enzyme substrate binding. In that way, the retrosynthesis approach produces an enzyme-annotated metabolic map containing all the pathways linking source metabolites to the target chemicals. Pathways are then enumerated and ranked to select the pathways that are best to engineer. Enumeration methods are based on steady-state analysis and network topology. The ranking function includes several criteria such as reaction feasibility, heterologous metabolites inhibitory effects and cytotoxicity, and chassis compatibility.

The CAD approach will be illustrated with the design and construction of 12 heterologous pathways and the characterization of intermediate reactions for the production of the flavonoid Pinocembrin in *E. coli* (an antioxidant precursors of several antimicrobials and anticancerous compounds). I will show how the CAD tools were used to design Pinocembrin biosynthetic pathways and to improve titer.

# Non Ribosomal Peptides : A monomeric puzzle

Yoann DUFRESNE[1], Valérie LECLÈRE[2], Philippe JACQUES[2], Laurent NOÉ[1] and Maude PUPIN[1]

[1] LIFL, UMR USTL/CNRS 8022, INRIA Lille-Nord Europe, 59655 Villeneuve d'Ascq, France
Yoann Dufresne : yoann.dufresne@etudiant.univ-lille1.fr, {maude.pupin, laurent.noe}@univ-lille1.fr

[2] ProBioGEM (UPRES EA 1026), Université Lille Nord de France, USTL, Polytech-Lille/IUTA, 59655 Villeneuve d'Ascq, France
valerie.leclere@univ-lille1.fr, philippe.jacques@polytech-lille.fr

**Abstract** *Nonribosomal peptides (NRPs) are increasingly studied because they harbor activities which can be exploited in various domains. They are often denoted as graphs illustrating their chemical structure, where the atoms are represented by nodes and the chemical bonds by arcs. Another possible representation is the monomeric structure. This structure, inspired by the biosynthetic pathway of these peptides, is effectuated by large enzymatic complexes which assemble together smaller compounds called monomers. Consequently, the nonribosomal peptides are composed of a great variety of monomers (more than 500 are known) including amino acids, lipids and carbohydrates. Likewise, nonpetidic bonds are formed between multiple monomers, producing peptides with cycles and/or branches. Thus, the monomeric structure is a graph formed by the monomers present in the peptide and their interlinking chemical bonds. Until now, there did not exist a tool allowing for the conversion between the atomic and monomeric structures. This article presents a novel algorithm capable of localising the monomers from a reference list in the chemical structures of peptides extracted from the Norine database. The algorithm is based on a heuristic that utilizes chemical information of NRPs. The preliminary results are encouraging, and should lead to further studies.*

**Keywords** nonribosomal peptides, chemical structures, graphs.

## Les peptides non-ribosomiques : un puzzle monomérique

**Résumé** *Les peptides non-ribosomiques (NRP) sont des molécules de plus en plus étudiées car elles présentent des activités ayant des applications principalement dans le domaine pharmaceutique. Elles sont souvent décrites par leur structure chimique, c'est-à-dire un graphe dont les nœuds sont des atomes et les arêtes les liaisons chimiques. Une autre représentation possible est la structure monomérique. Cette structure, inspirée de la voie de synthèse de ces peptides, est réalisé par de gros complexes enzymatiques qui assemblent les briques de base, appelées monomères. Ainsi, les peptides non-ribosomiques sont composés d'une grande variété de monomères (plus de 500 recensés jusqu'à présent) tels que des acides aminés, mais aussi des lipides ou des sucres. De plus, des liaisons non-peptidiques peuvent être formées entre certains monomères, ce qui produit des peptides contenant des cycles et/ou des branchements. La structure monomérique est donc le graphe formé par les monomères présents dans le peptide et les liaisons qui les relient. A l'heure actuelle, il n'existe pas d'outil permettant de convertir la structure chimique d'un peptide non-ribosomique en sa structure monomérique. Cet article présente un algorithme capable de localiser les monomères d'une liste de référence dans les structures chimiques des peptides de la base de données Norine. Il est basé sur une heuristique gloutonne qui utilise des connaissances sur la chimie des NRP. Les résultats préliminaires sont satisfaisants et devraient conduire à de nouvelles études.*

**Mots-clés** Peptides non-ribosomiques, structures chimiques, graphes.

## 1 Introduction

Les peptides non-ribosomiques sont synthétisés par certains micro-organismes (bactéries et fungi) et couvrent un large spectre d'activités biologiques [1]. Leur grande diversité de structures et de propriétés physico-chimiques est due à leur mode de synthèse, alternatif à la voie classique pour les peptides et protéines. Leur

synthèse est mise en œuvre par de grands complexes enzymatiques appelés synthétases peptidiques non-ribosomiques (abrévié NRPS en anglais).

Les peptides non-ribosomiques sont une source, encore sous-exploitée, de principes actifs dans l'industrie pharmaceutique, l'industrie cosmétique, le domaine des pesticides, des détergents et de la dépollution. Plusieurs de ces peptides sont déjà commercialisés tels que la pénicilline et d'autres antibiotiques, la cyclosporine qui réduit les risques de rejet de greffes ou l'actinomycine utilisée dans le traitement de certains cancers. Les étapes préliminaires à l'étude de la mise sur le marché de nouveaux principes actifs est la découverte de nouvelles molécules et l'étude de leur(s) activité(s). La réalisation expérimentale de ces étapes est coûteuse et demande une grande expertise. Il est cependant possible de réaliser une partie du travail via l'analyse bio-informatique des différentes sources de données à disposition afin de réduire le champ de criblage de manière efficace et ainsi diminuer fortement le temps d'investigation et les coûts engendrés.

À la différence des peptides classiques, les peptides non-ribosomiques (NRP) ne sont pas linéaires. Leur synthèse complexe vient modifier la structure linéaire pour ajouter des cycles et des branchements. Les briques de base (monomères) composant les NRP sont une deuxième source de différence avec les peptides classiques. Alors que les peptides classiques se basent sur les 20 acides aminés standards, les NRP s'appuient sur plus de 500 monomères différents dont environ 200 acides aminés, mais aussi des sucres et des lipides [2].

Dans les articles scientifiques et bases de données de petites molécules, les peptides non-ribosomiques sont décrits sous la forme de structures chimiques, c'est-à-dire un graphe dont les sommets sont les atomes et les arêtes sont les liaisons. Cependant, dans la base de données Norine et pour certaines utilisations comme la prédiction d'activité, il est intéressant d'avoir la structure monomérique, c'est-à-dire un graphe dont les sommets sont les monomères, chaque monomère étant composé de plusieurs atomes. La description sous forme de structure monomérique est inspirée du mode de synthèse des NRP. En effet, les synthétases sélectionnent spécifiquement les monomères puis les assemblent pour former un peptide. Ainsi, les monomères sont équivalents aux acides aminés ou aux nucléotides incorporés, respectivement, dans les protéines et les acides nucléiques. Pour l'instant, le passage de la structure chimique à la structure monomérique est réalisée manuellement et nécessite l'expertise de biochimistes qui découpent les monomères grâce à leurs connaissances concernant les NRP.

Dans cet article, nous allons présenter les travaux en cours au sein de l'équipe Bonsai du LIFL, en collaboration avec le laboratoire ProBioGEM, afin de permettre l'automatisation de cette tâche chronophage. Dans une première partie nous expliquerons pourquoi il est nécessaire d'avoir une approche différente de celle appliquée aux peptides classiques. Ensuite nous expliquerons les méthodes mises au point pour créer une première heuristique gloutonne qui a été testée sur la base de données Norine [3]. Enfin, nous parlerons des perspectives envisagées afin d'améliorer les performances de ce premier algorithme.

## 2 Objectifs

La méthode la plus utilisée pour connaître la séquence en acides aminés d'un peptide ribosomique est de traduire, à l'aide d'outils bio-informatiques, la séquence nucléique codant ce peptide. Il est également possible de faire du séquençage de peptides, mais cette technique expérimentale, lourde et coûteuse, est peu employée. Les NRP n'étant pas directement issus de la traduction des ARN, les outils bio-informatiques classiques ne sont pas transposables. Il est donc nécessaire de passer par la détermination expérimentale de la structure qui aboutit à une structure chimique. La décomposition d'un peptide en monomères revient à rechercher les graphes chimiques des monomères dans le graphe chimique du peptide que l'on cherche à décomposer.

La difficulté vient du fait que les monomères ne sont pas intégrés tels quels dans les NRP. En effet, à chaque formation d'une liaison entre deux monomères, certains atomes sont perdus (par exemple, dans le cas d'une liaison peptidique, un $H_2O$ est libéré avec la perte d'un $OH^-$ et d'un $H^+$). Et surtout, les liaisons effectuées ne sont pas toutes des liaisons peptidiques, ce qu'il faut prendre en compte.

## 3 Algorithme de conversion d'une structure chimique en structure monomérique

Dans un premier temps, nous avons développé une heuristique gloutonne basée sur une bibliothèque de fonctions pour la chimie appelée Openbabel [4]. Les molécules chimiques peuvent être considérées comme des graphes étiquetés et la bibliothèque intègre un format de représentation de ces graphes appelés SMILES (FIG. 1). Openbabel intègre également un outil de recherche de motifs dans des molécules que nous utilisons pour la recherche des monomères dans les peptides. La syntaxe appelée SMARTS [5] permet d'exprimer les motifs sous la forme d'expressions régulières de type SMILES.

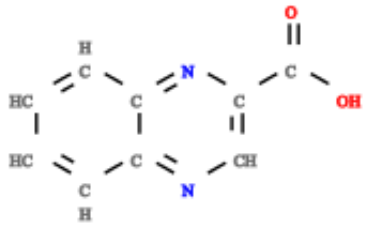



**Figure 1.** SMILES et structure chimique de la 2-carboxyquinoxaline

### 3.1 Radicaux et liaisons

Comme expliqué précédemment, en se liant les uns aux autres, les monomères se transforment. Pour les localiser dans les peptides, il est nécessaire de générer les différents *radicaux* : il s'agit de monomères qui ont perdus certains atomes lors de la formation d'une liaison. Afin d'inférer l'ensemble des radicaux possibles, nous avons étudié les liaisons formées dans les peptides de la base de données Norine.

La liaison peptidique est la plus fréquente. Elle est effectuée entre les groupements $NH_2$ et COOH (FIG. 2). Cette liaison peut parfois être légèrement modifiée. Par exemple, dans la proline, l'atome d'azote est dans un cycle et forme un groupement NH et non $NH_2$.

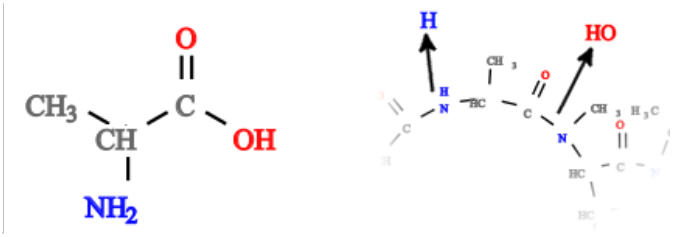


**Figure 2.** Alanine non liée - Alanine avec deux liaisons peptidiques

Certains acides aminés comme la cystéine possèdent un atome de soufre. Deux monomères portant un soufre peuvent former un pont disulfure (FIG. 3).

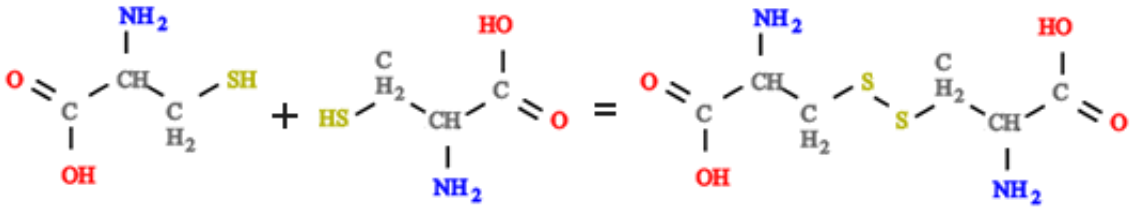


**Figure 3.** Formation d'un pont disulfure

Enfin, nous avons également observé des liaisons qui s'effectuent sur les cycles aromatiques. L'un des carbones de ces cycles peut perdre un hydrogène afin de former une liaison (FIG. 4).

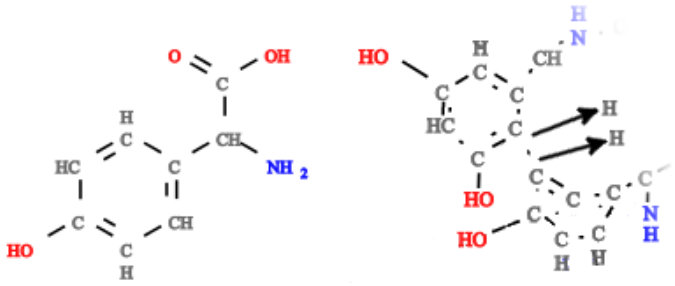


**Figure 4.** Hpg non lié - Deux Hpg liés dans la vancomicyne

Grâce à l'identification des liaisons présentes dans les NRP, nous sommes en mesure de construire des règles SMARTS pour localiser les atomes susceptibles d'être impliqués dans ces liaisons. Ainsi, nous pouvons générer tous les radicaux possibles pour chaque monomère (au total, 18030 radicaux pour les 531 monomères). Prenons l'exemple de la cystéine (FIG. 5). On peut distinguer sur la molécule trois groupements pouvant former une liaison (COOH, $NH_2$ et SH). Selon les peptides, ces liaisons ne sont pas toutes utilisées. En énumérant les cas possibles (liaisons uniques, paires et triplet), on peut générer 7 radicaux différents (FIG. 5).

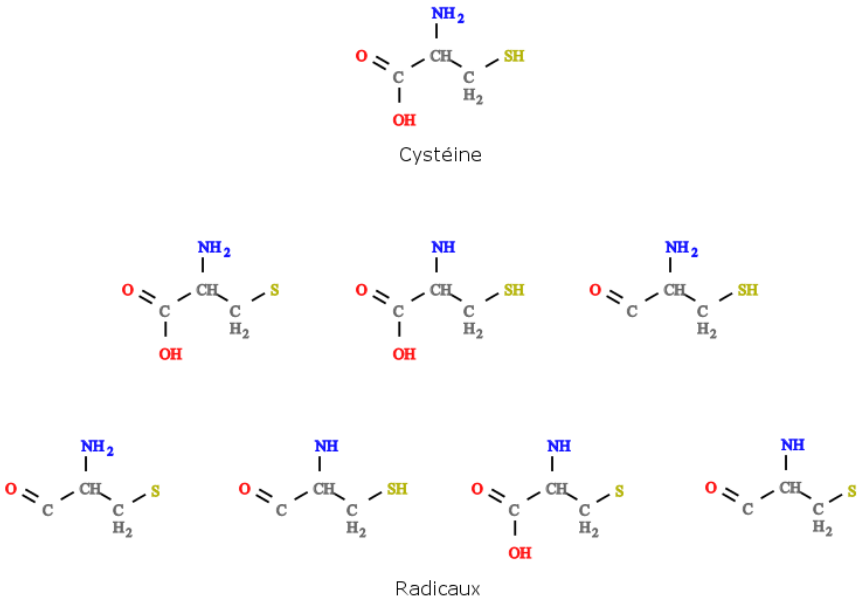



**Figure 5.** La Cystéine et ses radicaux générés

## 3.2 Heuristique gloutonne

Grâce aux règles précédemment générées et à l'API d'Openbabel, nous pouvons désormais rechercher les radicaux correspondant à l'ensemble des monomères au sein des peptides. Mais une question subsiste : "Quel monomère choisir lorsque deux radicaux différents peuvent être placés à un même endroit de la molécule ?"

Pour résoudre ce problème, nous avons choisi dans un premier temps d'appliquer une heuristique gloutonne de placement. On recherche séquentiellement les radicaux triés selon les critères définis ci-dessous. Dès que

l'un d'entre eux recouvre une partie du peptide, la partie couverte devient inaccessible pour les suivants (voir Algorithm 1).

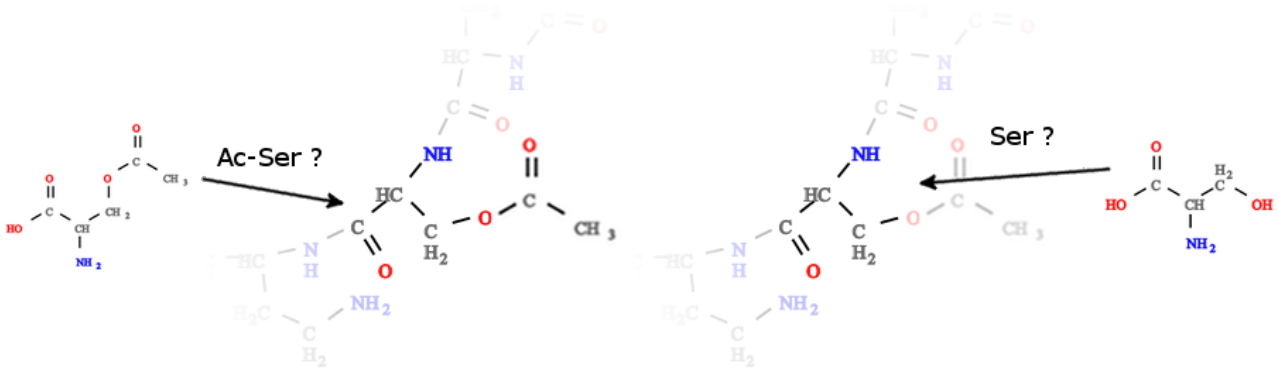



**Figure 6.** Exemple de peptide dans lequel plusieurs monomères peuvent être placés à un même endroit

Le premier critère discriminant choisi pour le tri est la taille (nombre d'atomes). La principale raison vient du fait que certains monomères et, a fortiori leurs radicaux, sont totalement inclus dans d'autres. Il est donc préférable, en règle générale, de placer les monomères les plus gros en premier pour que les petits ne prennent pas leur place. Par exemple, dans le peptide de la FIG. 6, il est possible de placer deux radicaux de deux monomères différents sur les mêmes atomes. De plus, certains radicaux d'un même monomère peuvent occuper le même emplacement, en couvrant plus ou moins d'atomes. Si le mauvais radical est choisi, il entrave des atomes. Ces atomes entravés peuvent, soit décaler le positionnement d'autres monomères, soit purement et simplement empêcher la pose d'autres monomères.

Un second critère de tri est appliqué pour discriminer les radicaux de même taille. Il s'agit de donner un score à chaque radical en fonction des liaisons qu'il peut former. Le poids de chaque liaison est déterminé en fonction de la fréquence de celle-ci dans les données. Par exemple, la liaison peptidique, étant majoritaire dans les NRP, a le poids le plus élevé. Le score d'un radical est alors défini comme la somme des poids des liaisons qu'il peut former. Ainsi, les radicaux peuvent être triés par taille puis score décroissants.

**Algorithm 1:** Heuristique gloutonne

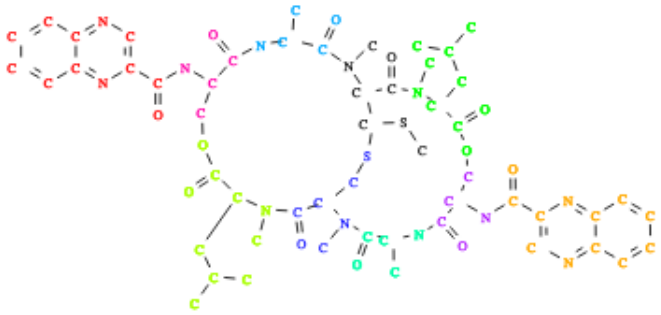
**Data**: Un peptide $P$ et une liste de monomères $M$
**Result**: Couverture de $P$ par des monomères $m$ tels que $m \in M$
Soit R une liste initialement vide de radicaux;
**forall the** $m \in M$ **do**
    Ajouter dans $R$ tous les radicaux possibles de $m$;
**end**
Trier $R$ selon la taille puis la pondération;
**forall the** $r \in R$ **do**
    **if** $r$ *match sur des atomes non couverte de* $P$ **then**
        Retenir le monomère $m$ ayant permis de générer $r$;
        Couvrir les atomes trouvés de $P$;
    **end**
**end**

## 3.3 Résultats

**3.3.1 Description de la sortie du programme.** Pour tester cette heuristique, nous avons utilisé les données provenant de la base de donnée Norine qui fait référence pour les NRP. Elle nous a permis d'obtenir les SMILES des 531 monomères qu'elle contient, ainsi que 204 peptides dont à la fois le SMILES et la structure

monomérique sont connus, parmi les 1200 peptides de la base. Dans cette section, nous allons présenter les résultats obtenus en comparant les prédictions du programme avec les annotations manuelles de Norine. La sortie du programme est disponible à l'adresse suivante :
`http://www.lifl.fr/~dufresne/norine/greedysplit/`

Voici comment interpréter ces pages : dans chacun des cadres, nous affichons les informations relatives à un peptide non-ribosomique. La première partie du cadre contient l'image générée par notre programme, c'est-à-dire le peptide coloré en fonction des monomères qui le recouvrent, ainsi que des indicateurs des performances de l'algorithme (FIG. 7). Le taux de couverture atomique (nombre d'atomes couverts / nombre total d'atomes dans le peptide), ainsi que le taux de couverture monomérique (nombre de monomères correctement trouvés / nombre de monomères dans le peptide) sont indiqués.



quinomycin B
Atomic coverage : 0.898
Monomeric coverage : 0.7

**Monomers:**

| found | wrong | not found | total |
|---|---|---|---|
| 7 | 2 | 3 | 10 |

**Figure 7.** Exemple de résultat pour un peptide

À la suite de cette première partie se trouvent trois listes (FIG. 8) représentant , respectivement :
- les monomères qui n'ont pas été trouvés dans le peptide alors qu'ils en font partie (FN) ;
- les monomères trouvés qui n'en font pas réellement partie (FP) ;
- les monomères trouvés qui sont vraiment présents dans le peptide (VP).

Pour aider le lecteur à localiser les monomères dans le peptide, une légende colorée est donnée.

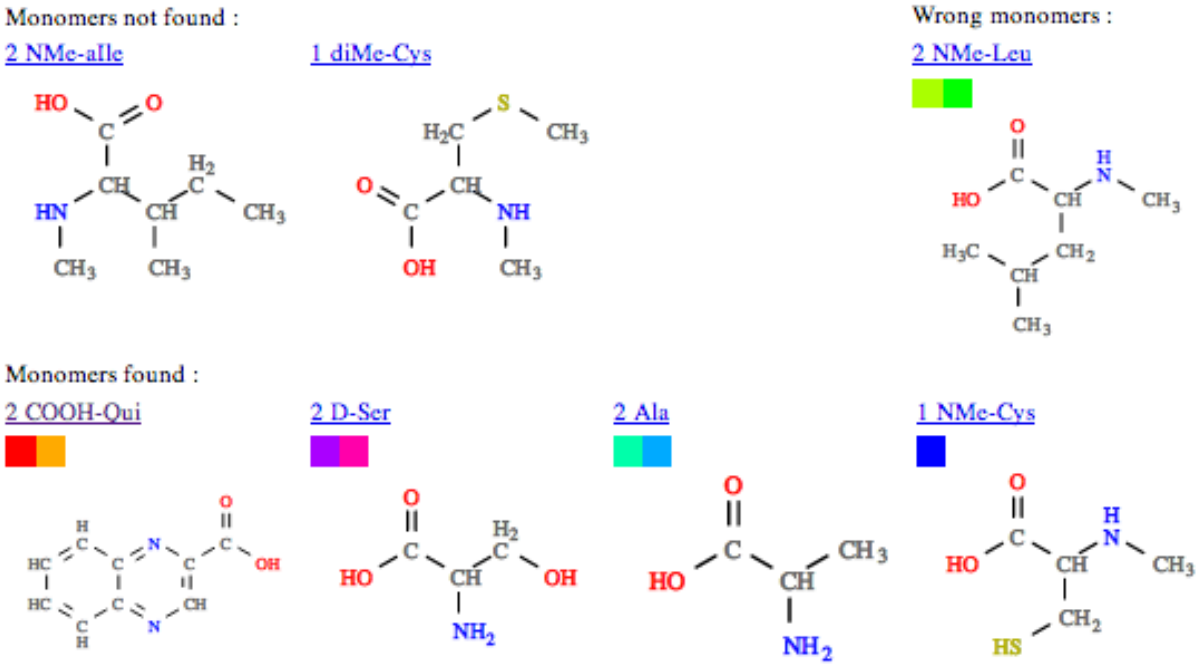



**Figure 8.** Listes de monomères

**3.3.2 Bilan chiffré.** Pour évaluer notre heuristique, regardons les couvertures atomiques et monomériques moyennes pour les 204 peptides extraits de la base de données Norine :

- Couverture atomique globale : 91.327%
- Couverture monomérique globale : 1552/1797 = 85.156%

Avec plus de 85% de monomères reconnus, nous pouvons dire que l'heuristique gloutonne donne de très bons résultats préliminaires. En effet, 1552 monomères ont été trouvés sur les 1797 possibles et seuls 188 ont été mal placés (voir TABLE 1). Concevoir un outil à partir de cette heuristique permettra ainsi aux biologistes d'avoir un assistant lors de l'annotation des peptides non-ribosomiques.

| VP ratio | FP ratio | FN ratio |
|---|---|---|
| 1552/1797 = 86,4% | 188/1797 = 10,5% | 245/1797 = 13,6% |

**Table 1.** Résultats

En examinant les peptides qui ont des monomères mal placés, nous avons observé que beaucoup d'erreurs viennent du fait que les monomères peuvent être observés sous différentes formes chimiques (tautomères, formes ionisées ou non, ...). La variation la plus fréquente est la tautomérie [6]. Il s'agit de doubles liaisons qui changent d'emplacement (FIG. 9). Ainsi, lorsque la double liaison n'est pas à la même position dans le SMILES du monomère et dans celui du peptide, nous ne sommes pas en mesure de localiser le monomère correctement. À cause de cette particularité, la recherche exacte de sous-graphes ne permet pas de trouver toutes les formes d'un monomère. Une équipe a défini 21 règles permettant d'énumérer les tautomères d'une molécule donnée [7]. Nous avons pour objectif d'intégrer ces règles dans notre dispositif, ce qui augmentera encore le nombre de radicaux générés à partir d'un monomère. De plus, il faudra prendre en compte le fait que la tautomérie peut se faire entre deux monomères adjacents dans un peptide.

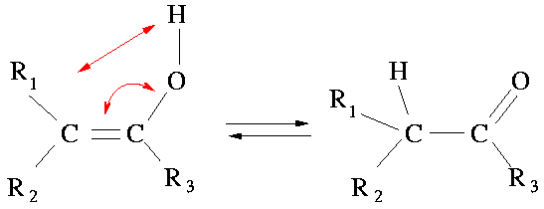



**Figure 9.** Exemple de tautomérisation

Nous avons également observé que d'autres erreurs proviennent de liaisons qui concernent des carbones à priori quelconques. Par exemple, la diMe-Cys de la quinomycin B (représentée en noir dans la FIG. 7) n'est pas trouvée car elle forme 3 liaisons avec d'autres monomères dont une directement sur un de ses carbones avec perte d'un hydrogène. Nous ne pouvons pas prendre en compte ces liaisons dans notre heuristique actuelle car la combinatoire serait trop élevée.

Enfin le dernier type d'erreurs ne provient pas du programme mais de la base de données Norine. Le test de notre programme nous a permis de mettre en évidence la présence d'erreurs dans certaines structures monomériques contenues dans Norine.

Pour conclure, nous pouvons affirmer que l'heuristique gloutonne donne des résultats de très bonne qualité puisque la plupart des monomères dont nous avons le SMILES sont retrouvés correctement dans les peptides. Il serait cependant intéressant de l'étendre en une heuristique plus souple permettant de réitérer le matching initial dans le but de l'améliorer (*hill-climbing*, algorithmes génétiques, ...).

## 4 Perspectives

La solution proposée ici repose sur une librairie n'offrant que la recherche exacte de motifs et non une recherche approchée. Le format SMARTS permet d'obtenir plusieurs variantes connues à l'avance d'une même molécule mais pas d'avoir des transformations inattendues. La recherche de SMARTS correspond à de l'isomorphisme de sous-graphe (*Subgraph Isomorphism*) [8]. Or, la catégorie d'algorithmes permettant la résolution

de ce problème n'est pas assez puissante pour résoudre les points bloquants qui sont les variations chimiques des monomères et les monomères inconnus.

Une solution qui permettrait de s'affranchir de la génération des radicaux serait de rechercher la plus grande sous partie commune entre le peptide étudié et le monomère recherché. On se ramène en fait à un autre problème connu dans le domaine de la théorie des graphes qui est le plus grand sous-graphe commun (*Maximum Common Subgraph*) [9]. Cette solution nous permettrait également de prendre en compte les éventuels monomères inconnus car il existe certainement des monomères qui ne sont pas encore dans Norine. Certains sont des dérivés de monomères connus et il serait intéréssant de pouvoir les inférer voir de découvrir dans certains cas par des approches comparatives de nouvelles formes de monomères. La version actuelle du programme ne permet pas de détecter automatiquement de nouveaux monomères. Elle peut cependant constituer une aide visuelle grâce à l'ajout de couleurs sur les images des peptides fournis en sortie du programme.

Enfin, l'ajout d'informations issues des connaissances biologiques concernant les peptides non-ribosomiques et leur mode de synthèse pourrait également permettre d'améliorer la qualité des résultats obtenus. Notamment, nous avons mené une étude statistique sur les données de Norine [2] qui a mis en évidence une différence de distribution entre les monomères présents chez les bactéries et ceux des fungi.

## Remerciements

Ce travail a été financé par Inria et le PPF Bioinformatique de l'Université Lille 1.

## Références

[1] G. Schoenafinger and M. A Marahiel, Nonribosomal peptides. in *Natural Products in Chemical Biology*, John Wiley & Sons, Inc, 2012.

[2] S. Caboche, V. Leclère, M. Pupin, G. Kucherov and P. Jacques. Diversity of monomers in nonribosomal peptides : towards the prediction of origin and biological activity. *Journal of bacteriology*, 192(19) :5143–5150, 2010.

[3] S. Caboche, M. Pupin, V. Leclère, A. Fontaine, P. Jacques and G. Kucherov. Norine : a database of nonribosomal peptides. *Nucleic Acids Research*, 26(D) :326–331, 2008.

[4] N. M O'Boyle, M. Banck, C. A James, C. Morley, T. Vandermeersch, G. R Hutchison. Open Babel : An open chemical toolbox. *Journal of Cheminformatics*, 3 :33, 2011

[5] Daylight Theory : SMARTS - http ://www.daylight.com/dayhtml/doc/theory/ theory.smarts.html.

[6] M.B. Smith, J. March. *Advanced Organic Chemistry (5th ed. ed.)*, Wiley Interscience, New York, 2001.

[7] M. Sitzmann, W.-D. Ihlenfeldt, M.C. Nicklaus. Tautomerism in large databases. *J Comput Aided Mol Des*, 24, 521–551, 2010.

[8] E. B. Krissinel and K. Henrick. Common subgraph isomorphism detection by backtracking search. *Software - Practice and Experience* 34 :591-607, 2004

[9] J. W. Raymond, P. Willet. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design*, 16 : 521–533, 2002

# Computational Protein Design in the Genomic Era

Thomas GAILLARD[1], Marcel SCHMIDT AM BUSCH[1,2], Anne LOPES[1], David MIGNON[1] and Thomas SIMONSON[1]

[1] Laboratoire de Biochimie, UMR7654 CNRS, Department of Biology, Ecole Polytechnique, 91128 Palaiseau, France
thomas.simonson@polytechnique.fr
[2] Institut fuer theoretische Physik, Johannes Kepler Universitaet Linz, Altenberger Strasse 69, 4040 Linz, Austria

**Abstract** *Millions of proteins are being identified every year by high throughput genome sequencing projects. Many others can potentially be created by protein engineering and design methods. Here, we review a method for computational protein design (CPD), which starts from a known protein and its 3D structure, and seeks to modify it by mutating some or all of the amino acid sidechains. The mutations are selected to provide stability, and possibly other properties, such as ligand binding. For each set of candidate mutations, the 3D structure is modelled, with an assumption of small, localized perturbations; in particular, we assume the backbone conformation does not change significantly. As in other CPD implementations, the structure is modelled using a classical, molecular mechanics approach along with a simple, implicit description of solvent. The method and selected results are described, which show that the designed sequences share important properties of natural proteins.*

**Keywords** protein design, inverse folding problem, directed evolution, molecular mechanics

## 1 Introduction: Structure Prediction on a Genomic Scale

Over the past decade, the genomes of several thousand organisms have been entirely "sequenced": the exact nucleotide sequence of the DNA that makes up their chromosome(s) has been experimentally determined [1]. These nucleotides (millions to billions, depending on the complexity of the organism) encode most of the molecules the cells need to produce, including their full complement of proteins. Proteins are the essential actors of the living cell: biochemical catalysts, motors, pumps, reading and interpreting the genetic message, directing the response to external signals or attacks. Humans, for example have a genome of about 3.4 billion nucleotides, including about 25,000 genes that code for proteins [2,3]. A protein is a polymer, or chain of amino acids, with a length usually between 100 and 1000 units. The amino acids are drawn from a small, natural "library" of 20 compounds [4]. The mapping that connects the nucleotide sequence and the amino acid sequence is known as the genetic code.

The challenge today is to determine the structure and biological function of all the known proteins [5,6,7]. Indeed, although the amino acid sequences of millions of proteins have been determined, most of their three-dimensional molecular structures are unknown. Yet the knowledge of these structures is essential to identify, understand, and possibly engineer or modify their biological functions. Predicting the three-dimensional structure from the amino acid sequence is the classic, "Protein Folding Problem", one of the most important problems in molecular biology today [5,8]. In the cell, the amino acid sequence of a protein uniquely directs it to "fold" into a specific, three-dimensional, molecular structure. In effect, the amino acid chain has a unique, preferred, three-dimensional arrangement, which corresponds to its lowest possible free energy [9,4]. It also has the ability to rapidly explore the available conformational space to find this preferred structure. The preferred structure is known as the "native" structure. The ability to fold rapidly into a unique, native structure is an essential and universal property of natural proteins. In contrast, a random, artificially constructed polymer of amino acids will almost certainly not "fold"; *i.e.*, it will not adopt a unique structure, but will divide its time among a large number of structures of comparable stability. Or worse: the sample will form an aggregate and precipitate in the form of a powder at the bottom of the test tube. Thus, over the course of millions of years of evolution, chance and natural selection have shaped modern protein sequences, building up a large, but very specific repertoire of viable sequences, capable of folding and performing a useful biological function.

In this paper, we describe a different approach. Instead of searching for the optimal conformation, or fold for a given amino acid sequence, we consider the inverse problem. For a given fold, we search for the best amino acid sequences [10,11]. With the Protein Folding Problem, we needed to search a vast conformational space. With the present, "Inverse Folding Problem", we need to search a completely different space: the space of amino acid sequences (of a given length). This brings us back to the "genomic space" from which we started out. We have solved the inverse folding problem for a collection of about 100 known protein structures [12,13]. This collection is made up of protein "domains", taken from the "Structural Classification of Proteins" (SCOP) database [14]. A protein "domain" is a structural unit, made of 50–300 amino acids, which is either a small protein, or a part of a larger protein. Domains represent an intermediate level of structural organization, since larger proteins are invariably built up from several distinct domains, and a protein domain can often fold into its specific structure by itself, even if it is removed from the rest of the protein to which it belongs [14,15]. An application of the inverse folding problem is the construction of new proteins, or "protein design". Among the sequences associated with a given protein domain, we can select those that are likely to perform a desired function, such as binding specifically to another protein, or catalyzing a particular chemical reaction. By selecting sequences that stabilize a given fold and, at the same time, are capable of performing a specific chemical or biological function, we perform molecular evolution in the computer. This technique for protein design is referred to as "Directed Evolution". Directed evolution has been successfully used in recent years to develop new biosensors, new catalysts, and to create completely new protein folds [16,17,18,19,20,21,22].

In the next sections, we describe our basic methodology for structure prediction and protein design, along with selected results obtained in the last few years [23,24,12,13]. The first ingredient of our method is a simplified, discrete description of the protein's conformational space when it is in the folded state. The second is a description of the unfolded state. The third is a classic, "molecular mechanics" description of proteins, which allows us to calculate the energy of any given sequence in any given conformation [25]. The fourth ingredient is a divide-and-conquer technique, where the necessary energy data are precomputed for the fold of interest, taking into account all possible sequences and sidechain arrangements. The overall complexity of this step is only quadratic with respect to the number of amino acids in the protein. The fifth is an algorithm and software for efficiently exploring sequence space, searching for the best sequences [19,23].

## 2 Protein Structure: the Importance of Being Discrete

A protein is a polymer of amino acids [4]. The amino acids are drawn from a natural library of 20 compounds. They share a common, backbone moiety, which is used to link them chemically in the polypeptide chain. The different amino acids are distinguished by their specific, sidechain moiety, which ranges in size and complexity. The backbone degrees of freedom give rise to the overall fold, while the sidechain degrees of freedom determine the local structure. In general, we are interested in the inverse folding problem: identifying the amino acids that stabilize a given protein fold. Therefore, we can assume the backbone degrees of freedom are fixed and concentrate on those of the sidechains [19,23].

The complexity of the sidechain conformational space remains formidable. It can be reduced to a manageable level, however, thanks to the "rotamer" concept, introduced by Janin et al [26] and exploited by Jay Ponder and Frederic Richards in their pioneering structure prediction work [11]. The sidechain geometries in proteins can be defined by a few torsional angles, corresponding to rotations of chemical groups around single chemical bonds. To explore the corresponding conformational space, it is very convenient to perform discrete steps along each torsional degree of freedom. This is especially well-suited to proteins, since in practice, some values of the torsion angles are much more probable than others. Although each amino acid type has typically 2–3 torsion angles, these adopt, on average, on the order of just ten preferred rotamers. Thus, using the discrete rotamer description has an enormous, simplifying effect on the protein's conformational space. For reviews of the preferred rotamers in protein structures, and for databases of preferred rotamers, see [27,28,29,30].

## 3 The Role of the Unfolded State

Protein stability is determined by a competition between the native, folded structure and an ensemble of unfolded structures. Indeed, the unfolded structure is not unique. When the protein deviates from its folded

structure, after an unusually violent collision with another molecule, for example, it will wander for a time among a large collection of less compact structures, before finding its way back to the most stable, native structure. In the language of thermodynamics, the existence of many different unfolded structures means that the unfolded state is stabilized by entropy. For a protein to function in the cell, it should spend most of its time in the folded state, since this is the 3D structure that is competent to perform the protein's function, be it catalysis of a biochemical reaction, transmission of a signal, or energy transduction. From thermodynamics, the time spent in the folded state increases exponentially as the folding free energy gets larger (more favorable). Therefore, to identify the most favorable sequences, we look for those that produce a large, favorable, free energy difference between the folded and the unfolded state.

Modelling the unfolded state is thus an essential ingredient of our structure prediction method. The structures of several thousand proteins in their folded state have been determined by Xray crystallography, as well as nuclear magnetic resonance and other techniques [31]. However, the unfolded state is very hard to characterize, because it is dynamic and poorly ordered. Therefore, following extensive earlier work, we adopt a very simple, empirical model of the unfolded state [32,19,23,33]. We simply assume the unfolded polypeptide chain is largely extended, so that the amino acid sidechains interact primarily with solvent, and only weakly with each other. This general organization is assumed not to depend much on the amino acid sequence (although the competition between folded and unfolded states does). Therefore, backbone interactions in the unfolded state will largely cancel when different amino acid sequences are compared.

## 4 Relating Structures and Energies: a Molecular Mechanics Description of Proteins

The most important computational models in use today for proteins are based on a "molecular mechanics" description. They represent the protein as a collection of spherical particles (the atoms), approximately incompressible, connected together by springs, each one bearing a small electric charge [25,34]. Solvent molecules can be described in the same way. To parameterize such a model for a large class of molecules like proteins takes several decades of researcher-years. Once in place, and despite its simplicity, a molecular mechanics model is a powerful tool to study the structure and stability of biomolecules.

A key element of the energy model is the description of the aqueous solvent which surrounds the protein. Indeed, the native structure is normally compact, or globular, and the protein folding reaction tends to segregate the less polar amino acids in the core of the structure, and the more polar ones at the surface. The less polar amino acid sidechains are made of alkane groups, which do not form very favorable interactions with water: they are said to be "hydrophobic". This segregation reduces the alkane–water interface, and the globular structure is stabilized by a "hydrophobic effect". Thus aqueous solvent plays an active role in driving the protein into its native structure, and structure prediction must take this into account. Yet it is much too expensive, for the applications below, to explicitly model thousands of water molecules, whose detailed behavior is not of interest *per se*. Therefore, most structure prediction methods rely on simplified descriptions of the aqueous solvent. In these descriptions, the solvent appears "implicitly", through its effect on the protein–protein interactions [35].

An essential property of the energy model just described is *pairwise additivity*: the energy takes the form of a sum of pairwise interactions between atoms or groups. It can be written:

$$U(r_1, r_2, \cdots, r_N) = \sum_i \sum_j U_{ij}(r_i, r_j), \tag{1}$$

where $i$, $j$ represent individual amino acids, $r_i$ is a vector that specifies the spatial positions of all the atoms of amino acid $i$, and $N$ is the number of amino acids in the protein (its length). Although the total energy $U$ (left) depends on the positions of all $N$ amino acids, it can be broken down into just $N^2$ terms (right), each of which has just "pairwise complexity", depending on just two amino acids $i$, $j$. This property makes possible the divide-and-conquer method described in the next section.

## 5 A Divide-and-Conquer Method for Protein Design

The divide-and-conquer approach described here was introduced by Mayo and coworkers [36]. It relies on two simplifications in the protein description: the pairwise energy function, explained in section 4, and the

simplified, discrete description of the protein conformational space, explained in section 2. With the protein backbone fixed and the rotamer approximation for the sidechains, we have just a few degrees of freedom for each amino acid. To find the optimal amino acid sequence and structure, we must consider just 20 amino acid types and $\sim$10 possible rotamers at each position. However, the size of the problem grows exponentially with the length of the protein chain, leading to a combinatorial explosion. For a small protein of 100 amino acids, for example, we have around $10^{100}$ structures for a single amino acid sequence. Considering all possible sequences (and a typical rotamer set), there are around $200^{100}$ structures: googols of googols.

Fortunately, we are using a pairwise energy function (Equation 1), and we can treat each amino acid pair separately. For a given amino acid $i$, and using a typical rotamer library, there are just $n$ = 200 possible combinations of amino acid types and rotamers. For a pair $ij$, there are then $n^2 = 200 \times 200$ combinations. In a protein of $N$ = 100 amino acids, there are $N^2$ such pairs. If we arrange the amino acids along the lines and columns of a two-dimensional table, we will need $N \times n$ lines and columns to tabulate all the energies, corresponding to all pairs and all combinations of amino acid types and rotamers. This table can be viewed as an energy matrix. It has $(Nn)^2$ = 400,000,000 elements in our example, which can be precalculated and stored. The complexity of this precalculation is only quadratic with respect to the protein length $N$, and the storage space needed will be only a few gigabytes. It is this precalculation that is the limiting step in our structure prediction and protein design methods. Once the energy matrix has been computed, the exploration of sequence and structure space can be done quickly and efficiently. The matrix calculation was implemented in the molecular mechanics program X-PLOR [37]. The scripts are available from our web site: biology.polytechnique.fr/biocomputing.

## 6 Exploring Sequence Space

For a given protein fold, defined here as the backbone structure, we want to identify the amino acid sequences that maximally stabilize the fold. Therefore, we need to solve an optimization problem in the space of sequences. An efficient search protocol, developed by Wernisch et al [32], intermixes sequence and structure changes, in a heuristic way. A "heuristic cycle" consists in the following steps. First, the amino acid type and rotamer at each position (or a large subset of positions) are randomized. Then, an iterative, steepest descent minimization is performed. The best amino acid type and rotamer are identified at the first position (given the current values at all the other positions); the best combination at the second position is identified, and so on. Each position is considered in turn, and multiple passes through the sequence are performed in this way, until no more improvement is obtained. This concludes the heuristic cycle; the final sequence represents a local optimum in sequence space. We emphasize again that the "best" sequence or rotamer means the one that maximizes the protein's stability; *i.e.*, the energy difference between the folded and unfolded states. After several 100,000 heuristic cycles, a representative set of good sequences is deemed to be obtained. The calculations are implemented in a C++ program, Proteus, initially adapted from the program Optimiser of Wernisch et al [32]. Proteus is available on our web site.

## 7 Selected Results: Performance of the Energy Function

To illustrate the quality of the computational model, we describe briefly some testing and parameter optimization. We apply our model to two generic problems. First, we predict sidechain positions for 29 proteins, given their backbone structures and amino acid sequence. This is the so-called sidechain reconstruction problem. Second, we predict the stability changes associated with a large set of point mutations in twelve different proteins.

For both problems, we optimized both the surface coefficients and the dielectric constant in the energy function (Section 4). A typical predicted structure is shown in Figure 1, and compared to the known, experimental, Xray structure. Most of the sidechains in the predicted structure have been correctly placed, and overlap nicely with the experimental sidechains. On average, for our test set of 29 proteins (about 3000 sidechains), 80% of the amino acids have their sidechains in the correct $\chi_1$ rotamer, similar to previous work with similar models. See [38] for a detailed description of our data.

For the stability changes, we considered 140 mutations in 12 proteins for which experimental stability measurements are available. After optimizing the surface coefficients and the dielectric constant (five adjustable

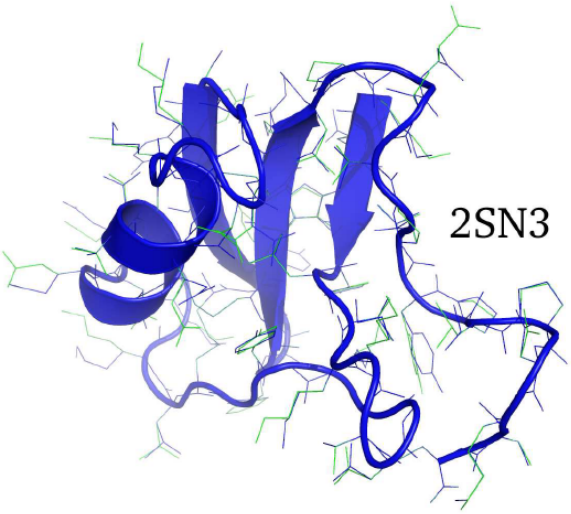



**Figure 1. Sidechain reconstruction** for the protein Staphylococcal nuclease. The protein backbone is shown in a simplified "ribbon" representation. Sidechains are shown as sticks, with the experimental positions (light blue) and computed positions (green) superimposed.

parameters), we could reproduce the experimental stability changes to within about 2 kcal/mol. This is comparable to other studies with models of this complexity [39]. However, this error level is still a bit high, given the exponential relation between errors in the energy and the time spent in the folded state (see above). Fortunately, the effects of the errors are alleviated by two factors. First, when a round of mutations is performed (at the beginning of each heuristic cycle, for example; see above), the energy errors for different mutated amino acids are random quantities that have a tendency to partially cancel each other. Second, an empirical correction is added to the energy function, such that our computed sequences tend to reproduce the correct, experimental abundancies of the different amino acids in natural proteins. This empirical correction is expected to reduce the error for the stability changes, so that the final mean error is somewhere between 0 and 2 kcal/mol. The quality of the corresponding sequences is illustrated below. See [24] for a detailed description of our data.

## 8 Generating Sequences for an Oncogenic Protein, the Src Homology 3 Domain

The Src Homology 3, or SH3 domain is one of the best-characterized members in the growing family of protein interaction modules. SH3 domains and their binding partners are abundant in species as different as yeast and *Homo sapiens*. Because of their involvement in protein–protein interactions, mutated forms of SH3 domains appear in several forms of cancer. Baker and coworkers carried out the complete redesign of the C-Src SH3 domain in 2003 [17]. Their predicted sequence failed to adopt the SH3 fold. Rather, it stayed unfolded, as revealed by biophysical experiments. In 2002, Wodak and coworkers [40] reported the complete redesign of 11 SH3 domains. Although the overall sequence identity between the computed and natural sequences was around 23%, the experimental sequences were poorly reproduced for surface regions of the proteins. Here, we describe our results on four completely redesigned SH3 domains: the SH3 domain of the Grb2 protein, that of the Vav protein, of the c-Crk protein and the cytoskeletal protein spectrin [12]. They include from 59 to 73 amino acids. Using X-PLOR, we computed the matrix elements for all pairs of amino acids. Protein sequences were then computed using the Proteus program. Proteus applies a heuristic procedure to search for the optimal amino acid and rotamer combination.

Inspection of the chosen 3D structures shows (see Figure 2) that a small beta strand [4] at the N-terminal part is followed by an extensive loop, which includes between 14 and 18 amino acids. The loop is then followed by additional four beta-strands, giving an ensemble of five antiparallel strands. Strands two and three, and strands three and four are connected by small turns, which comprise around five amino acids. Figure 2 compares experimental and computed sequences obtained for the Grb2 SH3 domain. The four upper sequences are experimental sequences from four different species: Rattus norvegicus, Pongo pygmaeus, Mouse, and Ze-

brafish. The fifth is the proto oncogene c-Crk (P38) (adapter molecule crk). Below, 25 computed sequences are listed [12]. These 25 proteins exhibit the highest similarity to the native sequences. Blosum matrix scores are used to score the sequences [41]. High scores indicate closely related sequences. Randomly chosen SH3 sequences, when compared to Grb2, have an average score of around 150. The depicted sequences score from 200 to 215. General speaking, Blosum distinguishes high favorable mutations (changes within one colored group), moderately favorable mutations (red to blue, orange to yellow), neutral mutations (orange to pink), and unfavorable mutations (blue or red to orange or yellow). Comparison of experimental and computed sequences reveals the conservation of many essential features of the native sequences, consistent with the high Blosum score of the calculated sequences.

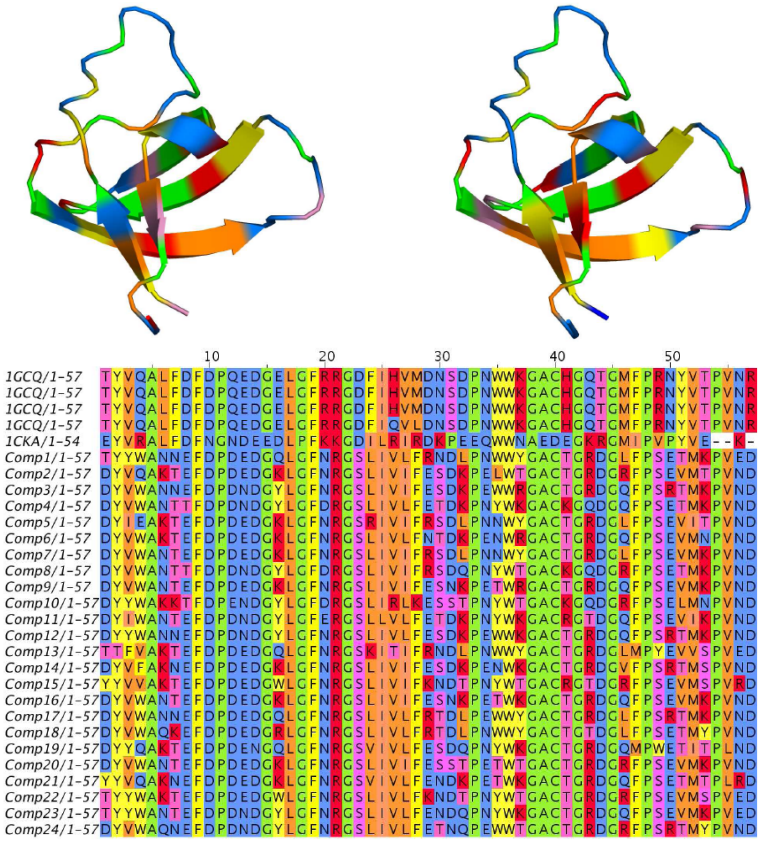



**Figure 2. Computed sequences. Top:** 3D Grb2 structures colored according to the experimental (left) and computed (right) sequences. The two color schemes are seen to agree qualitatively, indicating that the computed sequences reproduce the experimental pattern of amino acid types. **Bottom:** An "alignment" of experimental and computed sequences of the protein Grb2 (see text). The top four lines correspond to the experimental sequences of Grb2. The following lines correspond to representative computed sequences. To highlight the similarities (and differences) between the sequences, amino acids are colored according to the chemical properties of their sidechains (aliphatic, aromatic, polar, ionized, weakly-polar).

It is of interest to analyze the behavior in the core and at the surface, as mentioned. The two phenylalanines in positions nine and nineteen (marked as F in Figure 2) are core residues, and their two aromatic sidechains interact through $\pi$–$\pi$ stacking. This motif is reproduced in the computed sequences. The arginines in positions 21 and 22 are solvent-exposed. Here, we predict the mutation of the first arginine ("R") to asparagine ("N") and the conservation of the second arginine (giving a red column in Figure 2). The YV motif in positions 2–3 is conserved in the experimental sequences and occurs with a high abundancy in the computed sequences. From position 8 to 22 we predict sequences that are consistent with the experimental ones. The LF motif in position 6 and 7 is replaced by polar (Q,N) or ionic (K,K) amino acids, or by mixture of ionic amino acids and threonine (T). Inspection of the 3D structure shows that this LF motif is at the beginning part of a surface loop. Solvent exposed polar or ionized amino acids are energetically favorable in this type of position. For positions 24 and 25, we predict LI, compared to IL in the native c-Crk sequence. The WW motif at positions 35 and 36 of the

native sequence is conserved throughout the SH3 family. The computed sequences reproduce this motif in many cases.

An overall indicator of sequence quality is obtained by submitting the designed sequences to a standard fold recognition tool, such as the SUPERFAMILY library of Hidden Markov models [42,43]. We have done this for 25 SH3 domains, including the four above. When we select the 10,000 designed sequences with the lowest energies, 81% are correctly classified as SH3 domains by SUPERFAMILY. Results for 22 SH2 domains are equally good, with 83% of the low energy sequences correctly classified [12]. Similar results are obtained with several other protein families [13]. Finally, one designed SH3 sequence and five SH2 sequences were tested further by running molecular dynamics simulations for 4–6 nanoseconds with the protein fully solvated in a large box of water. The designed SH3 sequence led to RMS deviations from the backbone of the starting structure of about 1.7 Å, compared to 1.4 Å for the native sequence. Three SH2 sequences had deviations between 1.8 and 2.2 Å, compared to 1.4 Å for the native sequence. The SH3 sequence was produced experimentally and shown to fold (A. Sedano and P. Plateau, personal communication).

Overall, these and other published results [12,13] suggest that our method predicts amino acid sequences that reproduce important, native-like features. Hence, it has the potential to be useful for fold recognition, structure prediction and protein design.

## Acknowledgements

We thank the many volunteers who have participated in the Proteins@Home project and contributed computer cycles to this work. See biology.polytechnique.fr/proteinsathome for a complete list of participants. We thank the BOINC development community for testing the alpha version of Proteins@Home.

This article is adapted from: M. Schmidt am Busch, A. Lopes, D. Mignon, T. Gaillard & T. Simonson (2012) The Inverse Protein Folding Problem: Protein Design and Structure Prediction in the Genomic Era, In *Quantum Simulations of Materials and Biological Systems* (editors: J. Zeng, R. Zhang, H. Treutlein), Springer Verlag, New York.

## References

[1] R. F. Service. Gene sequencing: The race for the $1000 genome.

[2] E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[3] J. C. Venter et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[4] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing.

[5] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–96, 2001.

[6] O. Schueler-Furman, C. Wang, P. Bradley, K. Misura, and D. Baker. Progress in modeling of protein structures and interactions. *Science*, 310:638–642, 2005.

[7] T. Lengauer. *Bioinformatics: From Genomes to Drugs*. Wiley, New York, 2002.

[8] D. Lee, O. Redfern, and C. Orengo. Predicting protein function from sequence and structure. *Nature Rev. Molec. Cell Biol.*, 8:995–1005, 2007.

[9] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[10] D. Eisenberg. A problem for the theory of biological structure. *Nature*, 295:99–100, 1982.

[11] J. Ponder and F. M. Richards. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, 193:775–791, 1988.

[12] M. Schmidt am Busch, D. Mignon, and T. Simonson. Computational protein design as a tool for fold recognition. *Proteins*, 77:139–158, 2009.

[13] M. Schmidt am Busch, A. Sedano, and T. Simonson. Computational protein design: validation and possible relevance as a tool for homology searching and fold recognition. *PLoS One*, 5:e10410, 2010.

[14] A. Andreeva, D. Howorth, S. E. Brenner, J. J. Hubbard, C. Chothia, and A. G. Murzin. Scop database in 2004: refinements integrate structure and sequence family data. *Nucl. Acids Res.*, 32:D226–D229, 2004.

[15] F. Pearl et al. The cath domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucl. Acids Res.*, 33:D247–D251, 2005.

[16] G. A. Lazar, S. A. Marsall, J. J. Plecs, S. L. Mayo, and J. R. Desjarlais. Designing proteins for therapeutic applications. *Curr. Opin. Struct. Biol.*

[17] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302:1364–1368, 2003.

[18] L. L. Looger, M. A. Dwyer, J. J. Smith, and H. W. Hellinga. Computational design of receptor and sensor proteins with novel functions. *Nature*, 423:185–190, 2003.

[19] G. L. Butterfoss and B. Kuhlman. Computer-based design of novel protein structures. *Ann. Rev. Biophys. Biomolec. Struct.*, 35:49–65, 2006.

[20] S. M. Lippow and B. Tidor. Progress in computational protein design. *Curr. Opin. Biotech.*, 18:305–311, 2007.

[21] J. Pleiss. Protein design in synthetic biology. *Curr. Opin. Biotech.*, 22:611–617, 2011.

[22] I. Samish, J. M. Perez-Aguilar, and J. G. Saven. Theoretical and computational protein design. *Ann. Rev. Phys. Chem.*, 62:129–149, 2011.

[23] M. Schmidt am Busch, A. Lopes, D. Mignon, and T. Simonson. Computational protein design: software implementation, parameter optimization, and performance of a simple model. *J. Comp. Chem.*, 29:1092–1102, 2008.

[24] M. Schmidt am Busch, A. Lopes, N. Amara, C. Bathelt, and T. Simonson. Testing the coulomb/accessible surface area solvent model for protein stability, ligand binding, and protein design. *BMC Bioinformatics*, 9:148–163, 2008.

[25] A. D. Mackerell, Jr. Atomistic models and force fields. In O. Becker, A. Mackerell, Jr., B. Roux, and M. Watanabe, editors, *Computational Biochemistry and Biophysics*. Marcel Dekker, New York, 2001.

[26] J. Janin, S. Wodak, M. Levitt, and B. Maigret. Conformation of amino acid sidechains in proteins. *J. Mol. Biol.*, 125:357–386, 1978.

[27] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.*, 8:1267, 1991.

[28] R. L. Dunbrack and M. Karplus. Backbone-dependent rotamer library for proteins. application to sidechain prediction. *J. Mol. Biol.*, 230:543–574, 1993.

[29] R. L. Dunbrack and F. E. Cohen. Bayesian statistical analysis of protein sidechain rotamer preferences. *Prot. Sci.*, 6:1661–1681, 1997.

[30] R. L. Dunbrack. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.*, 12:431–440, 2002.

[31] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucl. Acids Res.*, 28:235–242, 2000.

[32] L. Wernisch, S. Héry, and S. Wodak. Automatic protein design with all atom force fields by exact and heuristic optimization. *J. Mol. Biol.*, 301:713–736, 2000.

[33] D. Seeliger and B. de Groot. Protein thermostability calculations using alchemical free energy simulations. *Biophys. J.*, 98:2309–2316, 2010.

[34] C. L. Brooks, M. Karplus, and M. Pettitt. Proteins: a theoretical perspective of dynamics, structure and thermodynamics. *Adv. Chem. Phys.*, 71:1–259, 1987.

[35] B. Roux and T. Simonson. Implicit solvent models. *Biophys. Chem.*, 78:1–20, 1999.

[36] B. I. Dahiyat and S. L. Mayo. De novo protein design: fully automated sequence selection. *Science*, 278:82–87, 1997.

[37] A. T. Brünger. *X-plor version 3.1, A System for X-ray crystallography and NMR*. Yale University Press, New Haven, 1992.

[38] A. Lopes, A. Aleksandrov, C. Bathelt, G. Archontis, and T. Simonson. Computational sidechain placement and protein mutagenesis with implicit solvent models. *Proteins*, 67:853–867, 2007.

[39] R. Guérois, J. E. Nielsen, and L. Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, 320:369–387, 2002.

[40] A. Jaramillo, L. Wernisch, S. Héry, and S. Wodak. Folding free energy function selects native-like protein sequences in the core but not on the surface. *Proc. Natl. Acad. Sci. USA*, 99:13554–13559, 2002.

[41] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89:10915–10919, 1992.

[42] M. Madera, C. Vogel, S. K. Kummerfeld, C. Chothia, and J. Gough. The superfamily database in 2004: additions and improvements. *Nucl. Acids Res.*, 32:D235–D239, 2004.

[43] D. Wilson, M. Madera, C. Vogel, C. Chothia, and J. Gough. The superfamily database in 2007: families and functions. *Nucl. Acids Res.*, 35:D308–D313, 2007.

# Session 5 : Réseaux d'interactions et structure des protéines

## Conférence invitée

CHRISTINE BRUN

TAGC, Marseille, France

**The predictive power of protein interaction networks**

Protein interaction networks are increasingly used to decipher the molecular bases of cellular functions because proteins involved in a molecular complex, a pathway or a biological process tend to interact with each other, thereby forming 'functional modules' (Hartwell et al. 1999). I will present the bioinformatics methods we developed to identify such modules, the benefits of the module-based approaches to predict function and the recent experimental validations we obtained.

**Reference**

- Hartwell, Hopfield, Leibler and Murray (1999) From molecular to modular cell biology. Nature 402: C47-52.

# MoMA-LigPath: a web server to simulate protein-ligand unbinding

Didier DEVAURS[1,2], Léa BOUARD[1,2], Marc VAISSET[1,2], Christophe ZANON[1,2], Ibrahim AL-BLUWI[1,2], Romain IEHL[1,2], Thierry SIMÉON[1,2] and Juan CORTÉS[1,2]

[1] CNRS, LAAS, 7 av du colonel Roche, F-31400 Toulouse, France
[2] Univ de Toulouse, LAAS, F-31400 Toulouse, France
{devaurs, lbouard, marc, zanon, ialbluwi, riehl, nic, jcortes}@laas.fr

**Abstract** *Protein-ligand interactions taking place far away from the active site, during ligand binding or release, may determine molecular specificity and activity. However, obtaining information about these interactions with experimental or computational methods remains difficult. The computational tool presented in this paper, MoMA-LigPath, is based on a mechanistic representation of the molecular system, considering partial flexibility, and on the application of a robotics-inspired algorithm to explore the conformational space. Such a purely geometric approach together with the efficiency of the exploration algorithm enables the simulation of ligand unbinding within very short computing time. Ligand unbinding pathways generated by MoMA-LigPath are a first approximation that can provide very useful information about protein-ligand interactions. When needed, this approximation can be subsequently refined and analyzed using state-of-the-art energy models and molecular modeling methods. MoMA-LigPath is available at http://moma.laas.fr.*

**Keywords** Protein-ligand interactions, molecular motions, robotics-inspired algorithms.

## 1 Introduction

In the past, experimental and computational approaches aimed at investigating protein-ligand interactions have mostly focused on the molecular complex, when the ligand is docked in the active site of the protein. However, an increasing amount of research shows that important interactions that may determine protein-ligand specificity and activity occur far away from the active site, during ligand binding or release (see for example [1,2,3,4]).

Despite impressive recent advances in structural biology techniques, obtaining accurate experimental data about protein-ligand interactions taking place far from the active site remains very difficult. Computational methods are therefore needed to better understand such interactions. However, simulating ligand (un)binding, particularly when the active site is deeply buried into the protein, is a challenging problem for current computational approaches. Some variants of molecular simulation methods have been devised specifically for that. In particular, Steered Molecular Dynamics (SMD) [5] and Random Acceleration Molecular Dynamics (RAMD) [6] have become popular techniques for the simulation of ligand (un)binding. Both methods introduce an artificial force in the molecular force field to enhance the ligand motion in a given direction. In SMD, this direction is defined by the user, whereas in RAMD, this direction is randomly chosen and iteratively modified after a given number of simulation steps if the ligand gets stuck. Although these methods have been shown to provide biologically relevant information, the artificial force introduced to accelerate the simulation may yield biased results about induced conformational changes, so that the interest of using an accurate molecular force field is partially lost. Monte-Carlo-based techniques have also been proposed for the study of ligand (un)binding and diffusion [7]. They perform a more computationally-efficient exploration of the conformational space than techniques based on molecular dynamics simulations, and do not require additional, artificial forces in the molecular force field to accelerate simulations. Nevertheless, all the aforementioned methods remain computationally expensive.

Our group has developed an original approach to simulate protein-ligand (un)binding, and other types of large-amplitude (long time-scale) molecular motions, at a very low computational cost [8,9,10]. It is based on a mechanistic representation of molecules, and on the use of methods inspired by robot motion planning algorithms to explore their conformational space. We have validated this approach in comparison to other computational methods and confronted it to experimental data. We have also successfully applied this approach to rational enzyme engineering [11,2]. Based on this novel methodology, we are developing a computer software called MoMA (for Molecular Motion Algorithms), which implements a collection of robotics-inspired algorithms for the simulation of molecular motions.

MoMA-LigPath is a web application built on the MoMA software. Starting from the model of a protein-ligand complex, MoMA-LigPath computes the ligand unbinding path from the active site to the protein surface. In the current version, flexibility is considered only for the ligand and the protein side-chains, and only geometric constraints are involved. Computing time for simulating ligand unbinding in such a simplified case usually ranges from some seconds to a few minutes. Thus, this simple version is particularly well suited to a web server. Even though they satisfy only geometric feasibility, the paths generated by MoMA-LigPath can provide interesting information to biologists and chemists. They can also serve as a first approximation that can be further refined using standard molecular modeling techniques. Note that more sophisticated algorithmic variants implemented in MoMA can consider protein backbone flexibility and energy models [10,12], at the expense of additional computational cost. These variants could be introduced in subsequent versions of the web server or distributed as binaries.

Next section gives explanations on the molecular models and the algorithm used in MoMA-LigPath. It also provides guidelines for users of the web server. After this, some results illustrating the capabilities of the method are presented.

## 2 Materials and Methods

### 2.1 Underlying methods

In recent years, algorithms originally developed for robot motion planning have been extended and applied to solve different problems in computational structural biology [13,14]. MoMA-LigPath applies one of these methods, based on the Manhattan-like RRT (ML-RRT) algorithm [15], originally proposed for disassembly path planning of complex objects with articulated parts. ML-RRT is an extension of the Rapidly-exploring Random Tree (RRT) algorithm [16], which iteratively constructs a tree that tends to rapidly expand on the search space with the aim of finding a feasible path between two given states. The main idea of ML-RRT is to divide variables (i.e. conformational parameters) into two groups, called active and passive variables, and to generate their motion in a decoupled manner. Active variables correspond to parts whose motions are essential for the disassembly task, whereas passive variables correspond to parts that should move only if they hinder the motions of other mobile parts (active or passive). The ML-RRT algorithm presents two main advantages compared with the basic RRT: First, ML-RRT can solve problems practically intractable with RRT. Second, it allows to automatically identify which parts have to move to find a solution to the disassembly problem.

The version of the ML-RRT algorithm currently implemented in MoMA-LigPath considers the ligand as an articulated mechanism to be disassembled from the protein. Besides, all the protein side-chains are also articulated with freely rotatable bonds. For the conformational exploration, the active variables are the parameters defining the pose (position and orientation) of a reference frame associated with the geometric center of the ligand, as well as the ligand bond torsions; the passive variables are the bond torsions of the protein side-chains. Collision avoidance between non-bonded atoms is the only feasibility condition considered for molecular motions. Further explanations about the approach are provided in previous publications [8,10].

### 2.2 Description of the web server

**2.2.1 Overview of the website.** The MoMA-LigPath website (http://moma.laas.fr) is structured into a set of pages. A *Home* page gives an overview of the website. *Demo* and *Help* pages provide guidance to users. A *Contact us* page is also included for particular requests. The main page, giving access to the application, is the

*Start a new job* page. The *My jobs* page, accessible only to registered users, contains a repository of previously submitted jobs. Finally, the *References* page lists some publications related to MoMA-LigPath. We kindly ask users to cite (at least) one of these publications if they use results provided by MoMA-LigPath in their work.

At this time, we highly recommend using Firefox, Safari or Chrome browsers. A correct display of the webpages is not guaranteed with Internet Explorer or other web browsers.

**2.2.2 Users and usage modes.** MoMA-LigPath is free and open to all users. Access as an anonymous user is possible. Nevertheless, we recommend frequent users to create an account. This will enable them to access a record of the jobs they submit to the server, and to be informed about results by email. Some privacy measures are enforced: the data and results of each user are not accessible to others. However, a completely secure communication pipe between each user and the server cannot be absolutely guaranteed.

Two usage modes are available to run jobs on the web server. The *simple* mode is the standard way of using MoMA-LigPath, and should be sufficient in most cases. The *advanced* mode enables the user to locally modify the flexibility of the molecules, and to access some internal parameters of the ML-RRT algorithm. Further explanations on these two usage modes are provided below.

**2.2.3 Input files and parameter settings.** The main input of MoMA-LigPath is a *.pdb* file containing the atomic coordinates of the protein-ligand complex. It can contain more than one protein and one ligand, as well as other molecules, such as structural waters or ions. However, only the last ligand described in this file is considered to be mobile in the current version.

When using the *advanced* mode, the user can locally modify the flexibility of molecules by uploading an additional input file with extension *.amc*, which is based on an internal file format. Given a *.pdb* file, a template *.amc* file can be generated via the web server, and subsequently modified with any text editor. The *.amc* file format is extremely simple. It contains blocks describing the flexibility of each molecule. For a protein, the block contains one line per residue. Each line contains the residue type, the residue identifier and two binary values defining the backbone and side-chain flexibility, respectively. A value of 0 means that all the dihedral angles of the backbone or the side-chain are fixed for that residue. A value of 1 means that they can freely rotate. Note that, in the current version of MoMA-LigPath, only the side-chains (and not the backbone) can be defined as flexible. By default, the flexibility of all side-chains is set to 1. Users can block side-chains by changing the values in their corresponding lines to 0. For a ligand, the block contains one line per rotatable bond. Each line contains the identifiers of the four atoms defining the dihedral angle, as well as the upper/lower bounds for the angle value. The user can remove dihedral angles or modify the bounds, which are set to $(-\pi, \pi]$ by default. Additional information about the *.amc* file format is provided in the template file header itself.

In both the *simple* and *advanced* modes, the user can optionally tune two basic parameters:

- % of van der Waals radii: This parameter is used for collision detection between non-bonded atoms. 80% is a reasonable default value, often used to check atom overlaps in other computational methods. A lower percentage may be necessary to find solutions to very constrained problems, which would require some flexibility of the protein backbone. In easy cases, this value can be increased to force the ligand to move along the medial axis of the exit channel.
- Number of paths: Up to 20 solutions can be requested for each job. Some variability can be observed in these solutions because of the random search performed by ML-RRT. Each solution path will be displayed individually in the results.

Additional parameters are accessible in the *advanced* mode:

- RRT expansion strategy: By default, MoMA-LigPath applies an RRT-Connect strategy to expand nodes during the construction of the search tree. This can be changed to a more basic RRT-Extend strategy, which generally produces shorter local paths, and is thus more computationally expensive. Nevertheless, the RRT-Extend strategy can be more efficient when solving very constrained problems. Please refer to the basic literature on the RRT algorithm [16] for further explanations on these strategies.

– Exit distance: The length of the paths to be computed by MoMA-LigPath can be specified by defining the distance (in Å) that has to be reached between the geometric centers of the ligand and the protein. When left unspecified, this distance is automatically computed by MoMA-LigPath, based on the molecule sizes.
– N fail max: This parameter determines the number of consecutive expansion failures after which a node in the search tree is considered "exhausted", and is no more selected for expansion during the ML-RRT construction. This heuristic is further explained in basic papers on ML-RRT [15]. The default value, 50, provides good results in general. This value can be increased for very constrained problems.

**2.2.4 Outputs and presentation of the results.** Registered users receive a notification email when a job terminates on the web server, and results appear on the *My jobs* page. For anonymous users, results are presented on a page having a unique URL.

The main output of MoMA-LigPath is a set of solution paths. Each solution, contained in a separate downloadable folder, is a sequence of *.pdb* files corresponding to intermediate conformations of the molecules along the ligand exit-path. The path is discretized in such a way that the maximum displacement of an atom of the ligand between two consecutive frames is approximately 0.5 Å. In addition, solution paths can be directly visualized on the results webpage by means of a Jmol applet [17].

For each solution, a file of *contacts* between the ligand and the protein is also generated. This is a text file containing a line for each pair of atoms in contact. A contact is detected if two atom overlap when their size is increased by 20% with respect to the size used for collision detection during the conformational exploration (e.g. if the path is computed considering 80% of van der Waals radii, contacts are identified for atoms overlaps at 100% of van der Waals radii). When the same contact appears for several conformations along the unbinding path, it is listed only for the first conformation. The number of the frame corresponding to this conformation in the sequence of *.pdb* files appears at the beginning of each line in the output file. Information on protein-ligand contacts may be interesting to identify important interactions. For instance, such information can help decision making for protein engineering [11,2].

In addition, an execution report provides information about possible errors in the input file and operations performed by the program. Information contained in this file is important to diagnose MoMA-LigPath failures. A frequent reason for failure is the presence of atom overlaps in the input *.pdb* file. In a pre-processing stage, MoMA-LigPath tries to remove such collisions by slightly perturbing the conformation of the concerned side-chains. If not all collisions are removed after a given number of iterations, the program execution stops. In this case, the user may solve the problem externally (e.g. by energy minimizing the model of the complex), or re-submit the job with a reduced % of van der Waals radii.

## 3 Results

This section briefly presents results obtained with MoMA-LigPath for the hexameric insulin-phenol complex, which is an interesting test system because of the likely existence of multiple pathways for phenol unbinding. The presented results only aim to illustrate the capabilities of the method, a further biological interpretation being out of the scope of this paper. More detailed explanations on the application of robotics-inspired algorithms to simulate protein-ligand unbinding in the context of enzyme engineering can be found in previous publications [11,2].

Insulin, in its monomeric, active form, is composed of two short peptide chains. In the presence of zinc ions, insulin monomers tend to associate, forming more stable hexameric structures [19]. Different conformational states of the insulin hexamer have been observed experimentally. Here, we consider the so-called $R_6$ state, which has a threefold symmetric, toroidal shape (see Fig. 1). This conformational state of the insulin hexamer is stabilized by bound phenolic molecules [20]. Understanding the mechanism of phenol unbinding is important because it is possibly involved in the conversion of the hexamer into the monomeric, active form of insulin [21]. Note that the study of the hexameric insulin-phenol complex and of the hexamer-monomer conversion are of interest in pharmacology for the treatment of type 1 diabetes.

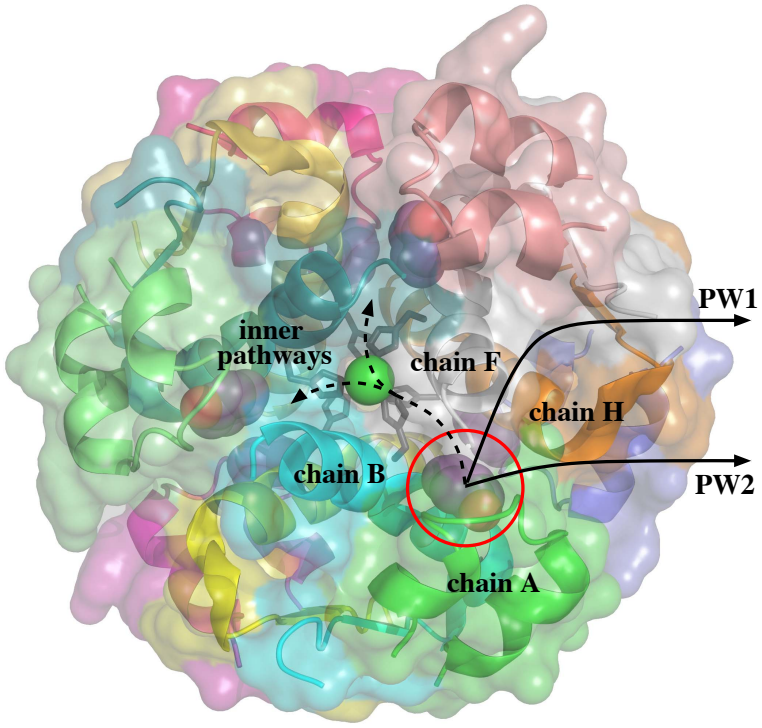



**Figure 1.** Structure of the $R_6$ hexameric insulin-phenol complex. The phenol molecule in the pocket between chains A, B, F and H can follow different unbinding pathways. The two most likely pathways are located at the interface of chains A, F and H. However, diffusion through the inner part of the hexamer is also geometrically feasible. Images of molecular models in this paper have been generated using PyMOL [18].

The structure of the $R_6$ insulin hexamer, determined by X-ray crystallography, is available in the Protein Data Bank [22], with PDB ID: 1ZNJ. The corresponding *.pdb* file was used as input for MoMA-LigPath. The $R_6$ insulin hexamer presents six hydrophobic pockets containing bound phenol molecules. MoMA-LigPath was applied to simulate the unbinding of one of them: the one located in the binding pocket between chains A, B, F and H. The other phenol molecules were kept in the input file, and considered to be static molecules. The side-chains of the six histidines interacting with the zinc ions where blocked by editing the input *.amc* file as explained in the previous section. The percentage of van der Waals radii used for collision detection between non-bonded atoms during the conformational exploration was set to 75% (the default value, 80%, had to be reduced because of overlapping atoms in the initial structure). Hydrogen atoms were not added. This is acceptable because the solutions provided by MoMA-LigPath are not expected to be an accurate representation of unbinding paths, but simply a first approximation.

MoMA-LigPath was run to simulate 20 phenol unbinding paths. To emphasize the computational efficiency of the method, we would like to mention that the average computing time for one solution was less than 10 seconds on a single processor. A significant variability in the solutions can be observed. In most simulations, the phenol molecule exits following pathways at the interface between chains A, F and H (see Fig. 1). Such paths can be clustered into two groups, which we refer to as pathway 1 (PW1) and pathway 2 (PW2), following the notation used in related work [21]. 25% of the paths (5 over 20) follow PW1. The ligand finds a passage between residues Ile10-A and His5-F, by inducing a significant conformational change of His5-F. Note that ring-flipping of His5 has been suggested by other computational and experimental studies on this system [21]. The ligand also induces the motion of the side-chain of Tyr16-H, which itself induces the motion of Tyr26-F. The conformations of other residue side-chains are also slightly perturbed by phenol unbinding following PW1. PW2 is the most frequent pathway observed in the solutions provided by MoMA-LigPath: 60% of the paths follow PW2. After a quick analysis of the solutions using a molecular viewer, one can observe that PW2 is the shortest and geometrically easiest pathway between the binding pocket and the protein surface, which explains the highest probability to obtain this type of solution. The ligand follows a narrow, partially open channel, and

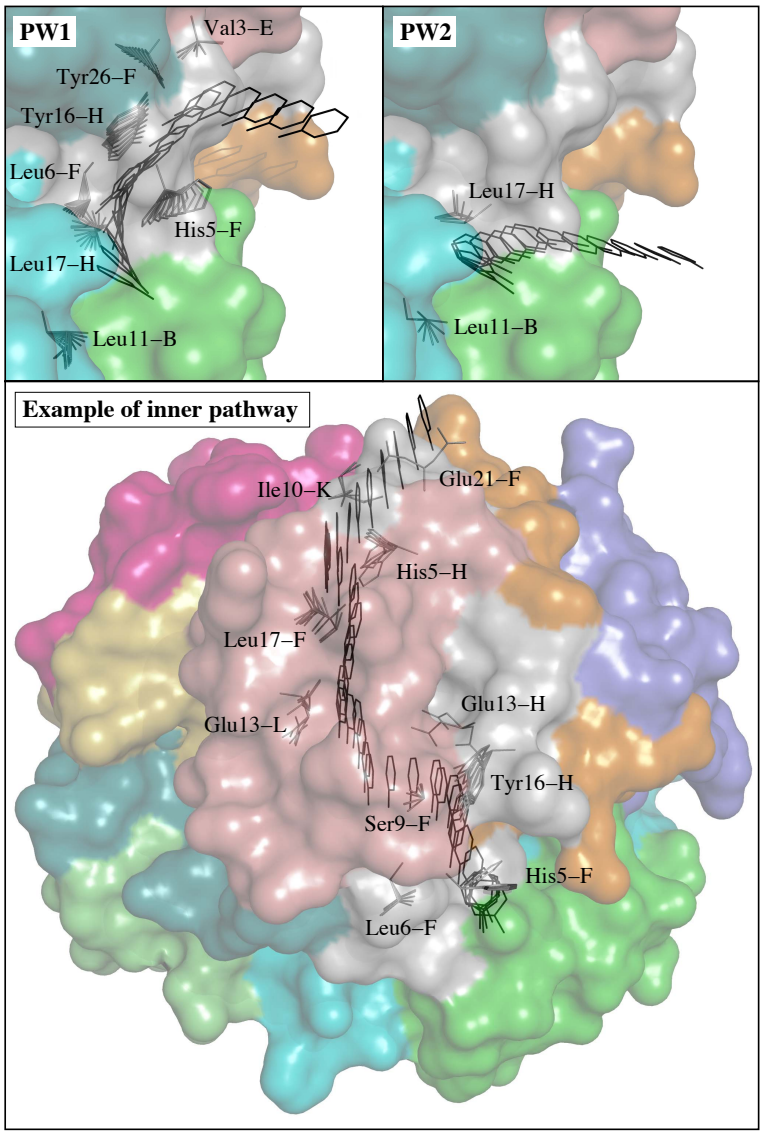



**Figure 2.** Different paths for phenol unbinding from the $R_6$ insulin hexamer obtained by MoMA-LigPath. The location of the phenol molecule and the conformations of moving side-chains are represented for some intermediate frames. The two images at the top correspond to paths following the most likely unbinding pathways: PW1 and PW2. The image at the bottom illustrates one of the pathways going through the inner part of the insulin hexamer.

induces slight conformational changes of only a few residues, mainly of Leu17-H. Remarkably, a combination of RAMD and SMD methods also pointed out PW2 as the most probable pathway for phenol unbinding [21]. Fig. 2 shows some intermediate frames of two solutions: one following PW1 and the other following PW2. The ligand and moving side-chains are represented. The figure also illustrates another type of pathway that was not reported in related work [21]: Surprisingly, in a few simulations (3 over 20), the ligand diffuses inside the insulin hexamer before finding an exit pathway. Indeed, the phenol molecule moves at the interface of insulin monomers toward the center of the hexamer, as indicated by the dashed line in Fig. 1, eventually finding exit channels that are the symmetric counterparts of PW1 and PW2. Note that the other phenol molecules, that are considered to be static in the simulations, do not obstruct these pathways. In two of the simulations, the ligand exits through a pathway similar to PW2 between chains F, H and K. One of such inner pathways is represented in Fig. 2. In another simulation, the ligand follows a pathway similar to PW1 between chains C, J and L. Exit through all the other pathways symmetric to PW1 and PW2 seems to be geometrically feasible. Note, however, that none of the 20 runs of MoMA-LigPath reported here, neither of the 100 additional runs performed in the same conditions, was able to find a third class of pathway (PW3) obtained by RAMD simulations as reported in [21]. PW3 is located at the interface between the two chains of the insulin monomer. It is a narrow corridor,

involving steric interactions of phenol with many residues. The fact that PW3 was not found by MoMA-LigPath means that it is geometrically unlikely, or even impossible, if the protein backbone does not deform. MoMA-LigPath was run again 20 times with a reduced atom size, namely 60% of van der Waals radii, to simplistically emulate slight fluctuations of the backbone. In this case, 10% of the solutions followed PW3. Nevertheless, since some of the results obtained with such a reduced atom size can be unrealistic, we prefer not to argue about the existence of this pathway type. More generally, a further analysis of the results presented in this section, considering energies, would be necessary to yield a more accurate model of phenol unbinding from the $R_6$ insulin hexamer.

## 4 Conclusion

This paper has presented MoMA-LigPath, which, to the best of our knowledge, is the first web application for the simulation of protein-ligand unbinding. MoMA-LigPath is based on a mechanistic representation of the molecular system, considering partial flexibility, and on the application of a robotics-inspired algorithm to explore the conformational space. The simplicity of the molecular model together with the efficiency of the exploration algorithm permit the simulation of protein-ligand unbinding within very short CPU time (generally, from some seconds to a few minutes), which enables the implementation as a web application.

Our aim with this web server is to provide easy accessibility to the methods we develop, which may interest a large scientific community. In particular, MoMA-LigPath can be an interesting tool for protein engineering, to help decision making for site-directed mutagenesis experiments. More generally, information on local protein-ligand interactions taking place far away from the active site may help understanding the overall molecular interaction mechanism. Finally, we would like to mention that MoMA-LigPath is not proposed as an alternative method, but rather as a complementary tool to be used in tandem with other computational and experimental methods.

## References

[1] Chaloupková, R., Sýkorová, J., Prokop, Z., Jesenská, A., Monincová, M., Pavlová, M., Tsuda, M., Nagata, Y., and Damborský, J. Modification of activity and specificity of haloalkane dehalogenase from Sphingomonas paucimobilis UT26 by engineering of its entrance tunnel. *J. Biol. Chem.*, 278(29):52622–52628, 2003.

[2] Lafaquière, V., Barbe, S., Puech-Guenot, S., Guieysse, D., Cortés, J., Monsan, P., Siméon, T., André, I., and Remaud-Siméon, M. Control of lipase enantioselectivity by engineering the substrate binding site and access channel. *ChemBioChem*, 10(17):2760–2771, 2009.

[3] Biedermannová, L., Prokop, Z., Gora, A., Chovancová, E., Kovács, M., Damborský, J., and Wade, R. C. A single mutation in a tunnel to the active site changes the mechanism and kinetics of product release in haloalkane dehalogenase LinB. *J. Biol. Chem.*, 287(34):29062–29074, 2012.

[4] Piechnick, R., Ritter, E., Hildebrand, P. W., Ernst, O. P., Scheerer, P., Hofmann, K. P., and Heck, M. Effect of channel mutations on the uptake and release of the retinal ligand in opsin. *Proc. Natl. Acad. Sci. U.S.A.*, 109(14):5247–5252, 2012.

[5] Isralewitz, B., Gao, M., and Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.*, 11(2):224–230, 2001.

[6] Lüdemann, S. K., Lounnas, V., and Wade, R. C. How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *J. Mol. Biol.*, 303(5):797–811, 2000.

[7] Borrelli, K. W., Vitalis, A., Alcantara, R., and Guallar, V. PELE: Protein energy landscape exploration. A novel Monte Carlo based technique. *J. Chem. Theory Comput.*, 1(6):1304–1311, 2005.

[8] Cortés, J., Siméon, T., Ruiz, V., Guieysse, D., Remaud-Siméon, M., and Tran, V. A path planning approach for computing large-amplitude motions of flexible molecules. *Bioinformatics*, 21(Suppl 1):i116–i125, 2005.

[9] Kirillova, S., Cortés, J., Stefaniu, A., and Siméon, T. An NMA-guided path planning approach for computing large-amplitude conformational changes in proteins. *Proteins*, 70(1):131–143, 2008.

[10] Cortés, J., Le, D. T., Iehl, R., and Siméon, T. Simulating ligand-induced conformational changes in proteins using a mechanical disassembly method. *Phys. Chem. Chem. Phys.*, 12(29):8268–8276, 2010.

[11] Guieysse, D., Cortés, J., Puech-Guenot, S., Barbe, S., Lafaquière, V., Monsan, P., Siméon, T., André, I., and Remaud-Siméon, M. A structure-controlled investigation of lipase enantioselectivity by a path-planning approach. *ChemBioChem*, 9(8):1308–1317, 2008.

[12] Jaillet, L., Corcho, F. J., Pérez, J. J., and Cortés, J. Randomized tree construction algorithm to explore energy landscapes. *J. Comput. Chem.*, 32(16):3464–3474, 2011.

[13] Gipson, B., Hsu, D., Kavraki, L. E., and Latombe, J.-C. Computational models of protein kinematics and dynamics: Beyond simulation. *Annu. Rev. Anal. Chem.*, 5:273–291, 2012.

[14] Al-Bluwi, I., Siméon, T., and Cortés, J. Motion planning algorithms for molecular simulations: A survey. *Comput. Sci. Rev.*, 6(4):125–143, 2012.

[15] Cortés, J., Jaillet, L., and Siméon, T. Disassembly path planning for complex articulated objects. *IEEE Transactions on Robotics*, 24(2):475–48, 2008.

[16] LaValle, S. M. and Kuffner, J. J. Rapidly-exploring random trees: Progress and prospects. In B.R. Donald, K.M. Lynch, and D. Rus, (eds.), *Algorithmic and Computational Robotics: New Directions*, A.K. Peters, Boston, pp. 293–308 , 2001.

[17] Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/.

[18] The PyMOL Molecular Graphics System, Version 1.5, Schrödinger, LLC.

[19] Dunn, M. F. Zinc-ligand interactions modulate assembly and stability of the insulin hexamer - a review. *Biometals*, 18(4):295–303, 2005.

[20] Berchtold, H. and Hilgenfeld, R. Binding of phenol to R6 insulin hexamers. *Biopolymers*, 51(2):165–172, 1999.

[21] Vashisth, H. and Abrams, C. F. Ligand escape pathways and (un)binding free energy calculations for the hexameric insulin-phenol complex. *Biophys. J.*, 95(9):4193–4204, 2008.

[22] Research Collaboratory for Structural Bioinformatics PDB: http://www.rcsb.org/pdb/.

# Session 6A : Analyse de séquences

# Whole genome re-sequencing : lessons from unmapped reads

Anaïs GOUIN[1,2], Fabrice LEGEAI[1,2], Pierre NOUHAUD[1], Guillaume RIZK[1], Jean-Christophe SIMON[1] and Claire LEMAITRE[2]

[1] INRA, UMR 1349, Institute of Genetics, Environment and Plant Protection, Le Rheu Cedex, France
{fabrice.legeai,pierre.nouhaud,jean-christophe.simon}@rennes.inra.fr
[2] INRIA Rennes - Bretagne Atlantique/IRISA, EPI GenScale, Rennes, France
{anais.gouin,claire.lemaitre}@inria.fr

**Abstract** *Unmapped reads are often discarded from the analysis of whole genome re-sequencing, while, opposingly, new biological information can be discovered from their analysis. In this paper, we investigated these reads from the re-sequencing data of thirty-three aphid genomes. The unmapped reads for each individual were retrieved from the results of the mapping of the sets of reads against the Acyrthosyphon Pisum reference genome, its mitochondrion genome and several known or putative symbiont genomes. These sets of unmapped reads were then cross-compared, this pointed out that a significant number of these sequences were conserved among individuals, especially when the latter are adapted to a same specific host plant, revealing that they may share crucial and functional material. Moreover, the analysis of the contigs resulting from the assemblies of the unmapped reads gathered by biotype allowed us to discover putative novel sequences absent from the reference genomes and highlighted the possible presence of other symbionts in the pea aphid genome whose existence were not known previously. As a conclusion, this study emphasizes that using a default strategy (e.g for the mapping) may lead to the loss of important information, and must be accompanied by specific analyses depending on the biological model.*

**Keywords** comparative genomics, NGS, re-sequencing, mapping, assembly, genomic variants, unmapped reads, aphid genomics

## 1 Introduction

NGS and whole genome re-sequencing is nowadays commonly used to identify genomic variants that are potentially involved in phenotypic variations, genetic diseases, adaptation or speciation in natural populations. Typically, the reads are mapped against a reference genome and the SNP and structural variant calling are based on the mapped reads [1,9]. Beside the usual caveats regarding unknown insertions or genomic contaminations, using these strategies on non-model organisms such as the pea aphid, may suffer from the poor quality of the reference genome and the uncompleteness of symbionts and organelles genomes. These problems produce a non negligeable fraction of unmapped reads, whose sequences are neglected in favor of the mapped in the further steps of the analysis, while they may contain useful information. This study describes our way to mine the unmapped reads in order to extract relevant biological knowledge, and lead to advice and recommendations for other re-sequencing projects.

We investigated this question in the context of a large scale re-sequencing project on the pea aphid complex. The pea aphid *Acyrthosiphon pisum* is a phytophagous insect feeding on more than 20 Fabaceae genera. This species forms a complex of sympatric populations, or biotypes, each specialized on one or a few legume species [12,14]. Peccoud *et al.* (2009) showed that these biotypes include at least eight partly reproductively isolated host races and three cryptic species, forming a gradient of specialization and differentiation potentially through ecological speciation [10]. In addition, the pea aphid is associated with an obligatory endosymbiont, Buchnera aphidicola, located in specialized cells called bacteriocytes and providing its host with essential amino acids. The pea aphid also harbors several facultative symbionts whose distribution is strongly correlated with plant specialization of their hosts and it has been posited that some of these symbionts could play a role in plant adaptation, although clear evidence is still lacking [8,13].

The study was carried out on thirty-three aphid re-sequenced genomes. The reads were mapped against the *Acyrthosiphon pisum* reference genome, its mitochondrion genome and several known (*Buchnera aphidicola*), or putative symbiont genomes. The quality of *Acyrthosiphon pisum* reference genome (530 Mb) composed of 23,924 scaffolds ([4]) is poor compared to those of model organisms, and some symbiont genomes sequences may not be well characterized for this species. As a result, an important part of the reads were not mapped. In this paper, we scrutinized these unmapped reads by performing cross-comparisons between the sets, assembling the reads by biotype and analysing the resulting contigs. We used new sophisticated tools such as *Minia* [3] and *Compareads* [7], and more classical ones such as Blast tools [2]. This analysis revealed that biological information are contained in the unmapped reads and we discovered putative novel sequences of the *Acyrthosiphon pisum* and new symbiont genomes.

## 2 Material and Methods

### 2.1 NGS data

Thirty-three aphid genomes were paired-end re-sequenced using the Illumina HiSeq 2000 instrument with around 15X coverage for each genome. The individuals belong to different populations referring as biotypes according to their adaptation to a specific host plant. We then have for this study, eleven biotypes composed of three individuals each. Reads are 100bp long, sequenced in pairs with a mean insert size of 250 bp. and 42.5 million read pairs were obtained on average for each individual.

### 2.2 Read mapping

The 100bp paired-end reads were mapped, using *Bowtie2* [5] with default parameters (up to 10 mismatches per read, or less if indels are present), to a set of several reference genomes simultaneously. This set is mainly composed of the official pea aphid reference genome, that is the *Acyrthosiphon pisum* reference genome [4], and its mitochondrion genome. It contains also the genome of the primary bacterial symbiont *Buchnera aphidicola*, and several known putative secondary symbiont genomes (*Candidatus Hamiltonella defensa* 5AT, *Candidatus Regiella insecticola* R5.15, *Rickettsia* sp. endosymbiont of *Ixodes scapularis*, *Rickettsiella grylli*, *Serratia symbiotica* str. Tucson, *Spiroplasma melliferum* KC3, *Wolbachia* sp. strain wRi). Various statistics about the quality of the mapping were carried out and we calculated for each individual the average coverage for each reference genome used. *Acyrthosiphon pisum* genome coverage is about 14.3X on average (min=10.6X and max=19.96X) and *Buchnera* genome is covered around 748.8X (min=138.08X and max=1509.03X). The coverage of the other symbiont genomes depends on the biotype and varies from 0X to 117.7X.

Fragments for which both reads of the pair did not map against the reference genomes were extracted from the BAM file (mapping result file) and used for the present study.

In order to check the quality of the unmapped reads, *Prinseq* [11] was used. Entire poly-N tail at the 3'-end was removed and low quality sequences were trimmed (if quality less than 20 over a window of 10 nucleotides). Only sequences of at least 66 nucleotides were kept for the analysis.

### 2.3 Comparison of unmapped reads

With the remaining reads, we created the sets of unmapped reads for each individual. *Compareads* [7] was used to compare the read content of these sets in a pairwise manner : it can find similar reads between two sets of reads without assembling them. A read of set A needs to share at least 2 non-overlapping kmers of size 33 with at least one read of set B to be considered similar. This gives two percentages of similarity between sets A and B : the percentage of reads of A similar to reads of B and vice versa. For all pairwise comparisons, a symetric similarity score was also provided, computed as follows : $\frac{AinterB+BinterA}{N_A+N_B}$, with $AinterB$ the number of reads in set A similar to reads in set B, $BinterA$ the number of reads in set B similar to reads in set A, $N_A$ and $N_B$ the total number of reads in sets A and B respectively.

The 33 samples were classified based on this similarity measure, using *R* software with the maximum distance for the distance matrix and the complete linkage method for hierarchical clustering. A heatmap of the classified samples was produced to display graphically the pairwise similarity scores.

### 2.4 Assembly

To get a sufficient coverage for assembly, we used the union of common unmapped reads between the 3 individuals of a same biotype, that is the reads present in at least one comparison between two individuals of a same biotype were all concatenated in one fastq file. The de novo assembler *Minia* [3] was used to assemble the common unmapped reads for each biotype. The following parameters were used : k-mers seen less than 3 times were filtered out and the size of the kmer for the De Bruijn graph was set to 31.

To calculate the contigs' coverage, the sets of unmapped reads were re-mapped against the contig sequences using *Bowtie2* [5] (default parameters) and the number of reads mapping each contigs were obtained with *Samtools* [6]. Following the average coverage observed from the reads mapped on the genome for one individual (15X), we considered that the contigs with a coverage from 20% to 60% were issued from the genome (*nuclear-like* coverage), and contigs with higher coverage may come from the symbionts (*symbiont-like* coverage) or repetitive sequences.

### 2.5 Comparison and analyses of contigs

*BLASTClust* was used for assessing whether contigs were similar between biotypes [2]. A match was retained between two sequences if they were 80% identical over at least 90% of each sequence length. To find the origin of the larger contigs, they were BLASTed against the aphid reference genomes (nuclear, mitochondrion and symbionts), and contigs with hits with an e-value below 1e-50 were considered as highly divergent region of the *A. pisum* or its symbiont genomes, i.e. contained reads that could not be mapped during the first mapping step. The remaining contigs were then BLASTed against the non-redundant nucleic database (NR) (blastx).

## 3 Results

### 3.1 A non negligeable fraction of reads do not map

For a given individual, there are between 0.6 and 7 Million pairs (mean = 1.3 Million) of reads that do not map on any of the reference genomes (nuclear genome, mitochondrion or known symbionts), that is both reads of the pair is unmapped. This constitutes an average of 3.7 % of the initial read sets, and most of these are of good quality, as shown in Fig. 1, since few reads were removed (about 17 % ) by quality trimming (see Methods).

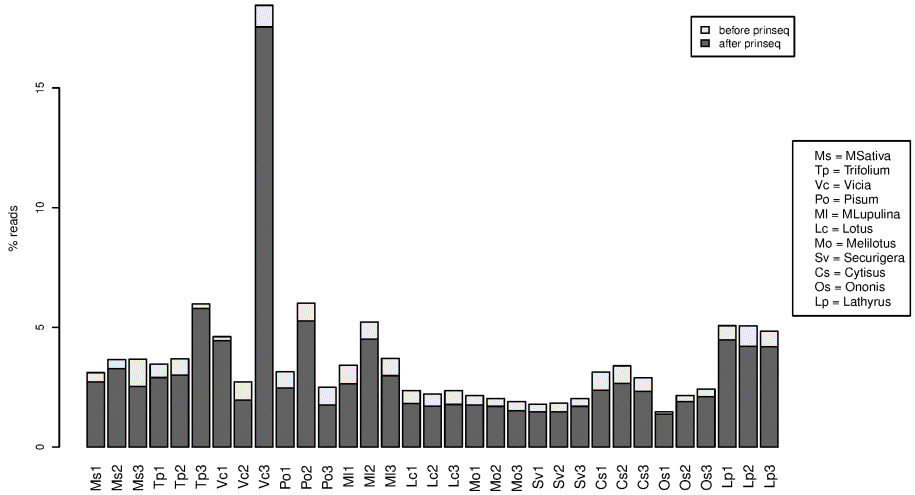



**Figure 1.** Percentage of unmapped reads (unmapped by pair) for each individual, after and before cleaning for quality. Individuals are grouped by biotype and sorted according to their known divergence with respect to the reference genome, the most divergent ones being at the right side of the figure.

We can also see in Fig. 1 that the fraction of unmapped reads varies between individuals. For some biotypes, the fraction of unmapped reads is very similar between individuals, suggesting a common cause of unmapping related to the biotype origin of the individuals. However the fraction of unmapped reads does not seem correlated to the divergence of the individuals (or biotypes) with respect to the reference genome.

This figure shows up that our mapping parameters were correct. Because, if we used a too stringent mapping parameters, we would have expected that the number of unmapped reads were correlated to the divergence of the biotypes and the reference genome, but this is not the case here.

## 3.2 There is some biologically meaningfull information in these reads

Each set of unmapped reads was compared to all other sets using *Compareads* [7]. Overall the 1056 (33x32) pairwise comparisons, the percentage of common reads between 2 individuals varies greatly, from 6 to 95% with an average value of 50% . The maximal of the average intersection percentage for each individual is 70% and each individual (except one) shares at least 50 % of its reads with one other individual. This strongly suggests that a large part of unmapped reads is not just random noise.

Interestingly there is a significant difference when comparing individuals of the same biotype (on average 70 % of common reads) versus individuals of different biotypes (48 %). This trend is confirmed by the hierarchical classification of individuals based on the pairwise similarity scores computed from the read set intersections (see Methods). Indeed, we can see in Fig. 2 that individuals belonging to the same biotype have comparable similarity profiles and are clustered together.

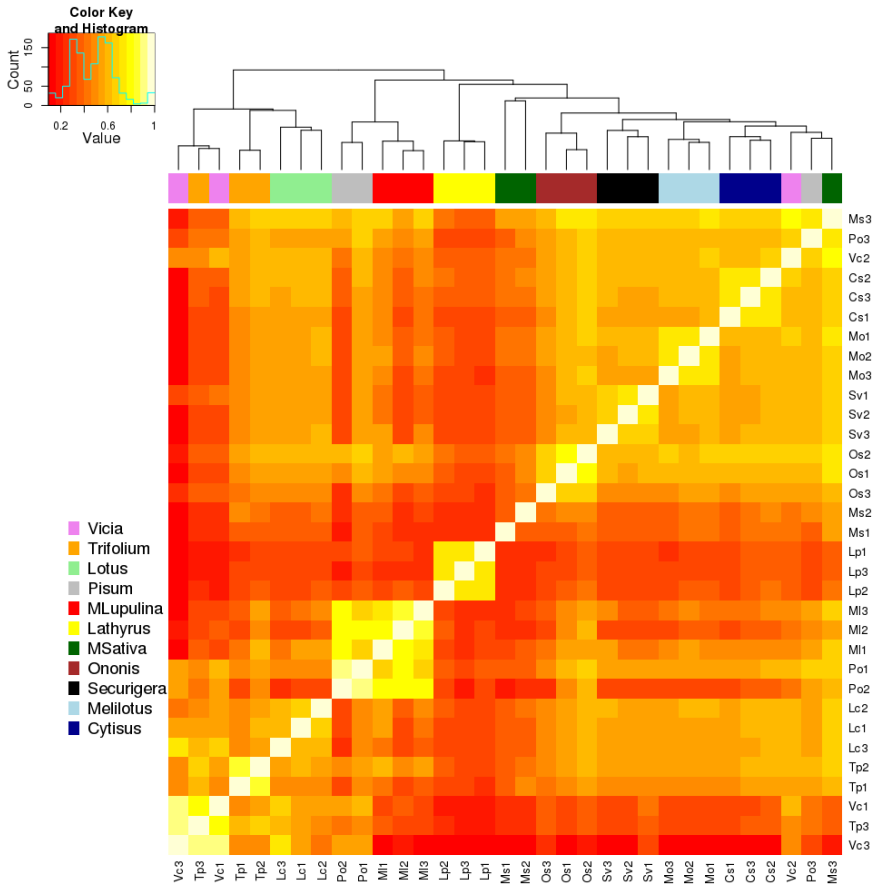



**Figure 2.** Hierarchical classification of the sets of unmapped reads. Each color below the tree corresponds to a biotype. Colors in the heatmap are function of the similarity score between two samples, from low similarity in red to high similarity in yellow.

In particular, one extreme and nice example is the *lathyrus* biotype, which is known to be the most divergent one (considered as a cryptic species) and which shows a very specific profile on the heatmap with strong similarity among this biotype (yellow group on Fig. 2) : a *lathyrus* individual shares on average 72 % of its reads with another *lathyrus* individual, whereas only 23 % with an individual of another biotype.

These results show that the sets of unmapped reads contain some sequence information able to discriminate the biotypes, and therefore unmapped read sets may contain valuable sequences for biological analyses. Indeed, the sequences they are consituted of may play a functional role in the speciation.

### 3.3 Where do these sequences come from ?

In order to get longer and more efficiently interpretable sequences, we assembled them conjointly by biotype, using the assembler *minia* [3]. Overall, 92.5 Mb of contig sequences ranging from 108 bp to 35.6 Kb were assembled. The average N50 is low (around 450 bp), but we get more than 12 000 contigs larger than 1 Kb (see Table in Fig. 3).

| Biotype | nbreads | all contigs | | | | Contigs > 500 bp | | | Contigs > 1 Kb | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | (M) | nb | Mb assbl | %reads | N50 | nb | Mb assbl | %reads | nb | Mb assbl | %reads |
| msativa | 3,68 | 20 202 | 6,04 | 40 | 402 | 2 748 | 2,61 | 27 | 739 | 1,24 | 16 |
| trifolium | 7,07 | 29 091 | 9,06 | 41 | 433 | 3 986 | 4,14 | 26 | 1 203 | 2,25 | 17 |
| vicia | 18,29 | 22 335 | 6,83 | 8 | 420 | 3 078 | 3,02 | 5 | 875 | 1,52 | 3 |
| pisum | 6,26 | 21 207 | 7,30 | 44 | 521 | 3 443 | 3,76 | 37 | 1 084 | 2,16 | 30 |
| mlupulina | 7,56 | 20 127 | 6,80 | 35 | 509 | 3 120 | 3,45 | 29 | 1 053 | 2,03 | 23 |
| lotus | 3,34 | 23 812 | 7,23 | 45 | 426 | 3 453 | 3,23 | 30 | 915 | 1,50 | 16 |
| melilotus | 3,68 | 23 120 | 6,99 | 41 | 428 | 3 305 | 3,17 | 27 | 927 | 1,53 | 15 |
| securigera | 2,96 | 22 465 | 6,79 | 48 | 421 | 3 216 | 3,06 | 31 | 915 | 1,48 | 17 |
| cytisus | 5,01 | 25 475 | 7,64 | 31 | 419 | 3 524 | 3,41 | 20 | 988 | 1,66 | 11 |
| ononis | 3,67 | 25 452 | 7,72 | 44 | 428 | 3 647 | 3,51 | 29 | 1 074 | 1,74 | 17 |
| lathyrus | 8,98 | 71 129 | 20,16 | 51 | 375 | 8 988 | 8,37 | 31 | 2 466 | 3,90 | 16 |

**Figure 3.** Contig statistics table. For each biotype, the number (nbreads) of unmapped reads used for the assembly is indicated along with several statistics describing the contigs for several length cut-offs, that is the number of obtained contigs (nb), their cumulative length (Mb assbl), the percentage of reads (%reads) that could be mapped on the contigs and the N50 value.

On average, 39 % of the reads could be remapped on the contigs. For contigs larger than 1 Kb, coverage varies greatly, with 70 % of them having a coverage around 45x suggesting that these contigs can originate from the pea aphid nuclear genome. On the other hand, there are 12 % of contigs with coverage larger than 60x which could originate from bacterial symbiont, mitochondrion, or repeated sequence.

But, we observe that some biotypes dissent from this trend : *mlupulina* and *pisum* have more contigs with a *symbiont-like* coverage, whereas *lathyrus* has almost all of its contigs with a *nuclear-like* coverage (see Fig. 4).

| Biotype | Coverage | | | | | | Blast match | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | nuclear-like | | | symbiont-like | | | on nuclear genome | | | on symbiont genomes | | |
| | % contigs | Total Kb | (%) | % contigs | Total Kb | (%) | % contigs | Total Kb | (%) | % contigs | Total Kb | (%) |
| msativa | 53 | 692 | 56 % | 13 | 192 | 15 % | 39 | 484 | 39 % | 16 | 144 | 19 % |
| pisum | 44 | 777 | 36 % | 33 | 1 032 | 48 % | 32 | 554 | 26 % | 43 | 955 | 60 % |
| mlupulina | 50 | 892 | 44 % | 36 | 940 | 46 % | 36 | 639 | 31 % | 50 | 884 | 63 % |
| vicia | 71 | 1 140 | 75 % | 5 | 90 | 6 % | 47 | 680 | 45 % | 0,2 | 1 | 0,1 % |
| lotus | 74 | 1 149 | 77 % | 5 | 87 | 6 % | 57 | 869 | 58 % | 0 | 0 | 0 % |
| securigera | 79 | 1 201 | 81 % | 5 | 71 | 5 % | 54 | 796 | 54 % | 0,2 | 1 | 0,2 % |
| trifolium | 52 | 1 215 | 54 % | 15 | 386 | 17 % | 35 | 702 | 31 % | 21 | 363 | 24 % |
| melilotus | 78 | 1 219 | 80 % | 5 | 74 | 5 % | 59 | 892 | 58 % | 0 | 0 | 0 % |
| ononis | 74 | 1 327 | 76 % | 5 | 88 | 5 % | 52 | 923 | 53 % | 0,2 | 1 | 0,2 % |
| cytisus | 78 | 1 336 | 81 % | 5 | 83 | 5 % | 56 | 949 | 57 % | 0 | 0 | 0 % |
| lathyrus | 88 | 3 463 | 89 % | 7 | 256 | 7 % | 62 | 2 438 | 63 % | 0 | 0 | 0 % |

**Figure 4.** Analyses of contigs larger than 1Kb in terms of read coverage and blast matches.

### 3.3.1 Too divergent aphid genomic sequences

On average, half of the contigs (of size at least 1Kb) have a significant blast hit with the nuclear reference genome. The contigs matching the reference nuclear genome have a mean coverage of 30x, consistent with a nuclear origin. Hence, these contigs likely originate from the nuclear genome and are assembled from reads that were too divergent to map in the first phase. When trying to map these reads with more sensitive parameters (*–very-sensitive* mode), less than 5% could be mapped on the reference genome, suggesting that this level of divergence may be not recovered with a pure-mapping approach. Consequently we observe that the *msativa* biotype, which is the same biotype as the reference genome, has the fewest contigs mapping to the genome than the others, and the more divergent *lathyrus* has the most summing to 2.4 Mb (see Fig. 4).

### 3.3.2 New bacterial symbiont

Some of the remaining contigs (not matching the aphid genome) have significant similarity with the symbiont genomes. But only 2 biotypes (*pisum* and *mlupulina*) have most of their contigs similar to symbionts (see Fig. 4). Accordingly, these contigs have a high coverage (more than 140x on average). They show similarity with mainly one symbiont genome : *Rickettsia* sp. endosymbiont of *Ixodes scapularis*. However, very few mapped on this genome, initially, with a coverage of only 3x for the *pisum* and *mlupulina* biotypes (and 0 for the others). This suggests that the chosen reference genome for *Rickettsia* were too distant from the actual aphid symbiont. When comparing these contigs to other *Rickettsia* species we could find a more closely related species *Rickettsia bellii*. But, the mapping of the reads on this new reference genome gave a lower coverage than the one observed from these contigs (around 40x combined for all three individuals of a biotype, instead of 140) suggesting that we still are missing a very close reference.

### 3.3.3 Analysis of the remaining contigs

We compared the contigs which neither mapped to the nuclear genome nor to the known symbionts, against the NR database, in order to identify regions which include highly divergent genes, or uncharacterized symbionts. Interestingly, beside various viral sequences, we found out sequences similar with many other insect genes, and few bacteria. We are currently proceeding to the refinement of the structures of these genes in order to facilitate their evolutional and functional annotations.

### 3.3.4 Common contigs, probably nuclear genome missing in the reference assembly

Some of these contigs are similar between several biotypes or even between all biotypes. We clustered them together using BlastClust. We obtained 60 clusters with contigs present in at least 8 biotypes and having a *nuclear-like* coverage. This represents 58 Kb of sequence, which certainly comes from the nuclear genome of the pea aphid.

Among these 60 clusters, 31 showed similarity with the reference genome and 23 with some aphid genes. These highly divergent genes may be considered as candidate genes for the host-plant adaptation. On the other hand, 29 contigs remained with no similarity with the reference genome. Hence, these latter contigs, representing 23.6 Kb, may be novel aphid sequences missing in the current reference assembly.

### 3.3.5 A special case of complex and highly repeated sequence

The third individual of the *vicia* biotype (vc3) shows atypical results compared to other individuals. First, it contains 5 times more unmapped reads than the average (more than 14 Million reads). Nevertheless, 90 % of these where found similar to another *vicia* individual (vc1). As the latter set contains only 4 Million reads, this high value suggests that the 14 Million reads set is highly redundant. Most of these reads were included in the reads set for the assembly phase, even so the obtained assembly was not longer than for the other biotypes and worse only 8 % of the reads could be mapped on the produced contigs.

Therefore with default parameters we were not able to assemble most of these reads. As it is not due to a low coverage since we showed that the set contained a lot of redundancy, we hypothesized that the reason of the

assembly failure was the high complexity of the assembly graph. We assessed this complexity by comparing the number of branching nodes versus non branching ones (branching node being small repeated sequences having different contexts). As a matter of fact, the *vicia* graph contains 10 % of branching nodes versus 4 % on average for the other biotypes. To reduce this complexity, we made a second assembly by setting the minimal kmer coverage parameter, usually used to filter out sequencing errors, to 10 (instead of 3), and obtained a larger assembly (23 Mb) using 6.7 Million reads (37 %). By setting the parameter to 100, we could still map 7 Million reads but only on 0.9 Mb. These latter sequences were thus highly covered, 700X on average.

These results show that the individual vc3 contains some sequences, not necessarly specific to this individual, but in very high coverage, and that are hard to assemble maybe because of numerous small sequence variants or sequencing errors difficult to filter out due to the high coverage. This results might be explained from a recent transposable element specifically active in the vicia biotype.

### 3.4 Discussion and conclusion

The analysis of unmapped reads from the thirty-three pea aphid re-sequenced genomes revealed that there are important biological information in these data, which are usually put aside. However their analysis is not trivial and we proposed a novel approach to rescue some of the lost information.

The direct pairwise comparisons of read sets, before assembly, enabled to find rapidly similar read sets and pinpoint atypical samples. Moreover, this enabled to choose combinations of samples to merge in order to achieve sufficient coverage for assembly. Indeed the coverage of each individual was too low to expect a good quality assembly. Nevertheless, selecting and merging only reads common to a biotype, may have prevented to find other interesting sequences specific to one genotype or to a combination of individuals of different biotypes. Therefore a more in-depth analysis of the pairwise comparisons followed by assembly of particular combinations of read sets would be interesting to conduct and may help to uncover unexpected links between individuals.

The assembly phase enables to obtain larger sequences that can be more efficiently analysed and compared to sequence databases. However, if bacterial sequences, such as the ones obtained from Rickettsia can be easily assembled and lead to large contigs, this is not the case for a majority of the unmapped reads. As shown by the Vicia example, some repeated and complex sequences need intensive parameter exploration or alternative assembly methods. Unfortunately, these peculiar reads may represent a large fraction of the unmmaped read sets as few of them could be remapped on the built contigs.

The final step of our approach was to align the contigs against sequence databases with less stringent similarity criteria (using the local aligner blast) than the one used during the first mapping step of our process. This enabled to determine the nuclear or symbiont origin of most of the larger contigs. For the symbiont origin, this revealed a wrong choice of reference genome and permitted to find a closer representative species. Without this analysis, we would have concluded from the first mapping that this symbiont was absent (or with a very low abundance) from all individuals.

Important biological implications can be extract regarding the contigs with nuclear origin. These are large regions either absent from the reference genome, or with sufficient divergence with the corresponding reference sequence so that each of the read pairs originating from it can not map. The latter explanation seems to be the most frequent in our dataset. It highlights the major drawback of classical comparative genomics approaches relying on a reference genome. The regions of the reference genome with important genomic divergence for some individuals will contain less mapped reads from these individuals and eventually few divergence will be detected, leading to an erroneous interpretation. Hence this mapping issue could lead to the loss of valuable biological information or biases in the variation analyses depending on the divergence of the individuals to the reference genome. One could think that lowering the mapping threshold to account for high levels of divergence could be a solution. However, this may lead to false positive mappings and more importantly will be too demanding in time and computer resources for huge high throughput sequence data sets. Here, our approach helps to recover those divergent regions, and in our specific case, the forthcoming analysis of those

rescued sequences will help to say if they are genomic regions harboring candidate genes involved in host-plant adaptation.

Finally, some of our contigs are novel sequences (or divergent duplicates), absent from the reference genome and could also harbor some important biological or evolutionary traits. We are currently investigating this with a different and complementary approach : the detection and assembly of inserted sequences using the whole read sets (not only the unmapped ones). We plan to compare the contigs obtained by both approaches and if some match are found, we will get an additional information of the location of the novel inserted sequences in the reference genome.

## Acknowledgements

This work was supported by ANR Blanc SPECIAPHID (ANR 11 BSV7 005 01) and by the Région Bretagne (Adaptro).

## References

[1] 1000 Genomes Project Consortium, R. M. Durbin, G. R. Abecasis, D. L. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 2010.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[3] R. Chikhi and G. Rizk. Space-efficient and exact de bruijn graph representation based on a bloom filter. In *WABI*, volume 7534 of *Lecture Notes in Computer Science*, pages 236–248. Springer, 2012.

[4] International Aphid Genomics Consortium. Genome sequence of the pea aphid acyrthosiphon pisum. *PLoS Biol*, 8(2):e1000313, Feb 2010.

[5] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat Methods*, 9(4):357–359, Apr 2012.

[6] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[7] N. Maillet, C. Lemaitre, R. Chikhi, D. Lavenier, and P. Peterlongo. Compareads: comparing huge metagenomic experiments. *BMC Bioinformatics*, 13(Suppl 19):S10, Dec 2012.

[8] A. H. C. McLean, M. van Asch, J. Ferrari, and H. C. J. Godfray. Effects of bacterial secondary symbionts on host plant use in pea aphids. *Proceedings of The Royal Society B: Biological sciences*, 278(1706):760–766, 2011.

[9] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and snp calling from next-generation sequencing data. *Nat Rev Genet*, 12(6):443–451, Jun 2011.

[10] J. Peccoud, A. Ollivier, M. Plantegenest, and J.-C. Simon. A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. *Proc Natl Acad Sci U S A*, 106(18):7495–7500, May 2009.

[11] R. Schmieder and R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, Mar 2011.

[12] J.-C. Simon, S. Carré, M. Boutin, N. Prunier–Leterme, B. Sabater–Muñoz, A. Latorre, and R. Bournoville. Host–based divergence in populations of the pea aphid: insights from nuclear markers and the prevalence of facultative symbionts. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1525):1703–1712, 2003.

[13] T. Tsuchida, R. Koga, and T. Fukatsu. Host plant specialization governed by facultative symbiont. *Science*, 303:1989, 2004.

[14] S. Via. Specialized host plant performance of pea aphid clones is not altered by experience. *Ecology*, 72(4):1420–1427, 1991.

# Follow-up of minimal residual disease in leukemia using high-throughput sequencing

Mathieu Giraud [⋆1], Mikaël Salson[⋆1], Marc Duez[1], Jean-Stéphane Varré[1], Céline Villenet[2], Sabine Quief[5], Aurélie Caillault[3], Nathalie Grardel[3], Christophe Roumier[3,4], Claude Preudhomme[3,4], Martin Figeac[2]

[1] Laboratoire d'Informatique Fondamentale de Lille (LIFL, UMR CNRS 8022, Univ. Lille) et Inria Lille, France
[2] Plateforme de génomique fonctionnelle et structurelle – Univ. Lille 2, IFR-114, Lille, France
[3] Laboratoire d'hématologie, CHRU Lille, France
[4] Inserm U-837, Institut de recherche sur le cancer, Lille, France
[5] Institut pour la recherche sur le cancer de Lille (IRCL), Lille, France

{mathieu.giraud,mikael.salson}@lifl.fr

**Abstract** *High-throughput sequencing offers new perspectives for leukemia follow-up. We propose an algorithm able to cope with millions of sequences, identifying and quantifying V(D)J recombinations in lymphocytes. It is now possible to follow a multi-clonal lymphocytic or lymphoblastic population and to measure its reaction to the cure. The proposed method is implemented in an open-source software called Vidjil, and tested on several samples of a patient suffering acute lymphoblastic leukemia.*

**Keywords** Sequence analysis, high-throughput sequencing, immunology, V(D)J recombinations, hematology, leukemia, minimal residual disease follow-up.

## Suivi de la leucémie résiduelle par séquençage haut-débit

**Résumé** *Le séquençage à haut débit offre de nouvelles perspectives pour le suivi de la leucémie. Nous proposons un algorithme pouvant traiter des millions de séquences, capable de différencier les réarrangements V(D)J qui s'opèrent au sein des lymphocytes, et de les quantifier. Il est désormais possible de suivre une population clonale de lymphocytes ou de lymphoblastes au cours du temps et, en pathologie, de mesurer sa réaction au traitement. La méthode proposée est implémentée dans un logiciel open-source appelé Vidjil, et testée sur plusieurs échantillons d'un patient atteint de leucémie aiguë lymphoblastique.*

**Mots-clés** Analyse de séquences, séquençage haut-débit, immunologie, réarrangements V(D)J, hématologie, leucémie, suivi de maladie résiduelle.

## 1 Introduction

Le répertoire immunitaire contient une grande diversité de lymphocytes, capable de déclencher de nombreuses réactions immunitaires. Une contribution importante de cette diversité (estimée à $10^{12}$) est due aux réarrangements des « gènes VDJ » (voir Figure 1) ainsi qu'à des mutations additionnelles au niveau des jonctions [16]. Les réarrangements VDJ surviennent uniquement pour la production des chaînes lourdes des lymphocytes (IgH) et les chaînes $\beta$ et $\delta$ des lymphocytes T (TR$\beta$ et $\delta$). Les réarrangements VJ sont propres aux chaînes légères des lymphocytes B (Ig$\kappa$ et Ig$\lambda$), et aux chaînes $\alpha$ et $\gamma$ des lymphocytes T (TR$\alpha$ et $\gamma$).

Analyser les réarrangements V(D)J permet de mieux comprendre la diversité du répertoire d'un individu, et de suivre l'évolution de pathologies. Dans le cas de la leucémie aiguë lymphoblastique (ALL, Acute Lymphoblastic Leukemia), hémopathie touchant principalement les enfants, il est possible d'identifier au diagnostic chez plus de 90 % des patients un réarrangement V(D)J qui pourra servir de marqueur des clones malins. Ce marqueur est utilisé pour le suivi de patients pour quantifier la maladie résiduelle [6,16]. Pour une meilleure efficacité du suivi, il faudrait détecter les réarrangements clonaux à de très faibles concentrations ($10^{-5}$ et en dessous). Les techniques actuelles (Biomed-2 [16] et Q-PCR) peuvent difficilement atteindre un tel niveau et, surtout, ne sont pas efficaces pour suivre l'évolution d'une population variée de clones [11].

⋆. Co-premier auteurs

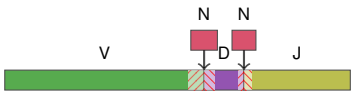



**Figure 1.** Un réarrangement VDJ comprend trois régions V, D et J de la ligne germinale, éventuellement tronquées ou mutées. De plus, quelques nucléotides de N-diversité peuvent être insérés au niveau des jonctions entre les trois régions. Un réarrangement VJ ne contient pas de région D. Les gènes V font typiquement 210 à 310 pb, les gènes D 10 à 30 pb, et les gènes J 35 à 65 pb.

*Analyse de réarrangements V(D)J.* Certains logiciels ont été conçus pour la segmentation V(D)J, c'est-à-dire l'identification des régions V, D et J dans une séquence donnée. Cette identification s'obtient par comparaison de la séquence testée aux séquences de référence des lignées germinales, comparaison qui peut se faire par programmation dynamique (JoinSolver [15], V-Quest [5], HighV-Quest [2]), éventuellement précédée d'heuristiques à base de graines (igBlast [1]), par des modèles HMM (iHMMune-align [7], SoDA2 [13]) ou par des méthodes basées sur le maximum de vraisemblance (VDJSolver [14]). Une évaluation de certaines de ces méthodes a été publiée dans [10].

Depuis 2009, plusieurs articles ont étudié le répertoire V(D)J en utilisant le séquençage à haut débit (SHD), chez le poisson zèbre [18] ou chez l'humain pour le suivi de la leucémie [4,8,12] ou pour explorer la diversité du répertoire lymphocytaire [3,17]. Le séquençage à haut débit (SHD) permettra à terme une précision plus fine que les méthodes conventionnelles et surtout un suivi de toute une population clonale. À l'heure actuelle, détecter l'émergence de nouveaux clones en cours de traitement ne se fait pas aisément avec les méthodes conventionnelles.

Les outils existants de mapping ou de clustering pour le traitement des données de SHD ne sont pas exploitables en l'état. Ils ne sont généralement pas capables de traiter des réarrangements avec des mutations dans une région d'intérêt de quelques dizaines de nucléotides. De plus, les résultats attendus pour une analyse de clonalité ne sont pas les segmentations V(D)J brutes de milliers ou de millions de reads, mais leur regroupement en clones. Les outils habituels de clustérisation ne peuvent pas être utilisés, car des séquences avec très peu de différences peuvent provenir de différents clones, particulièrement si ces différences sont dans la zone de N-diversité.

## 2 Algorithme proposé

Les outils mentionnés ci-dessus ont été principalement conçus pour étudier quelques séquences avec des réarrangements V(D)J, et certains d'entre eux prennent plusieurs heures pour traiter des millions de séquences. De plus, étant donné un séquençage haut-débit, avoir une analyse complète de chaque séquence n'est pas nécessairement pertinent: l'important est de quantifier les clones les plus présents, que cela soit pour une analyse du répertoire individuel d'un patient ou pour le suivi de maladie résiduelle.

Nous présentons donc ici une méthode rapide d'analyse des réarrangements V(D)J sur des données de séquençage haut-débit. Le but final de notre méthode n'est pas l'analyse fine de millions de séquences (la segmentation) mais la quantification de l'abondance relative des différents clones séquencés. Notre méthode comprend deux étapes :

– **Une prédiction heuristique de "$w$-jonctions"**, qui sont des régions de longueurs $w$ (par défaut 40) contenant la vraie jonction V(D)J. Après avoir construit un index de tous les "$k$-mots" (mots de longueur $k$) spécifiques aux segments V et J de référence, chaque read est balayée (voir Fig. 2): une segmentation sur le brin sens est détectée si l'on trouve des $k$-mots spécifiques aux segments V suivi de $k$-mots spécifiques aux segments J. Une jonction de longueur $w$ (par défaut 40) est extraite, centrée sur cette prédiction (le milieu entre la prédiction de fin du V et de début du J).

– **Une quantification des clones, s'appuyant sur les $w$-jonctions.** En l'absence d'erreurs de séquençage, toutes les $w$-jonctions extraites pour un même clone sont strictement identiques, même si elles ne sont pas exactement centrée sur la vraie jonction V(D)J à cause de mutations par rapport aux séquences de référence. Une séquence représentative du clone est choisie parmi tous les reads. Cette séquence peut ensuite être segmentée dans ses composants V(D)J en utilisant n'importe quel outil d'analyse existant [2,5,7,13,14,15]. Nous avons aussi implémenté un segmenteur simple alignant la séquence contre

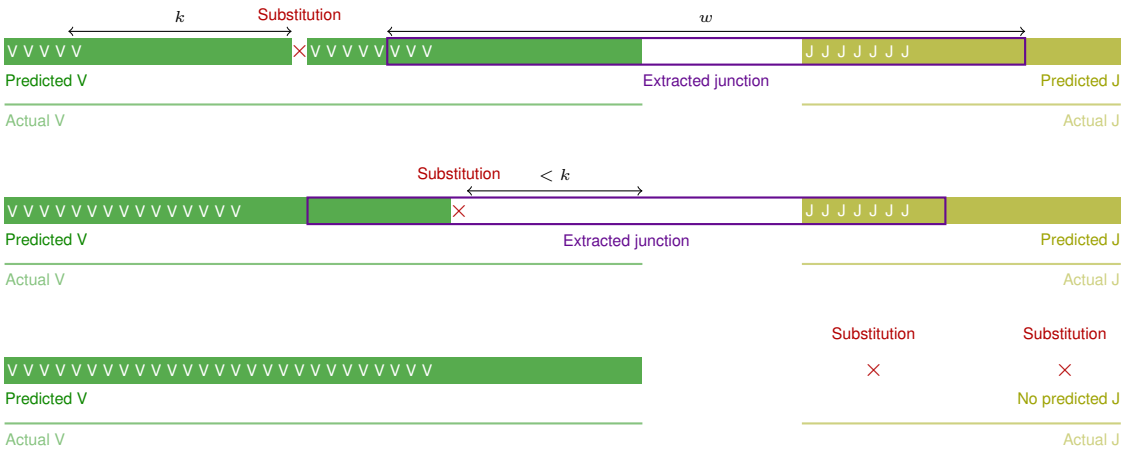



**Figure 2.** Heuristique détectant une $w$-jonction sur le brin sens dans des réarrangements VJ. La détection sur le brin anti-sens est faite de manière similaire, et la détection dans des réarrangements VDJ s'appuie aussi sur les segments V et J. Les étiquettes $V$ et $J$ indiquent le début de $k$-mots présents dans l'index. (Haut.) L'heuristique prédit correctement le centre de la jonction. Il y a une mutation (ou erreur de séquençage), représentée par $\times$, située loin en amont de la vraie jonction. (Milieu.) Une mutation ou une erreur dans les $k$ derniers nucléotides du segment $V$ mène à une erreur de prédiction dans le segment $V$ : en effet aucun $k$-mer exact ne peut être trouvé à la fin du V, le dernier $k$-mer trouvé dans le V est celui qui précède l'erreur. Cependant, comme de grandes valeurs de $w$ sont utilisées, la vraie jonction est toujours contenue dans la jonction prédite. (Bas.) Quand il y a trop d'erreurs (comparé à la taille des segments de référence), l'heuristique ne permet pas de prédire une jonction. Cela peut arriver particulièrement dans le segment $J$ qui est plus court.

les séquences de référence par programmation dynamique, en temps $O(\ell r)$ où $\ell$ est la longueur de la séquence et $r$ la longueur totale des séquences de référence.

La prédiction des jonctions étant en temps linéaire, la méthode pourra faire face à des dizaines de millions de reads : seul un petit nombre de jonctions est étudié en détail pour les clones les plus présents. La méthode ne nécessite pas non plus de grandes quantités de mémoire : pour des petites valeurs de $k$ (par défaut 14), l'index est directement stocké sous forme de table de taille $O(4^k)$.

**Regroupement de jonctions.** Les erreurs de séquençage (Fig. 2, milieu) se traduisent par des $w$-jonctions différentes, qui peuvent être regroupées: il est possible de le faire de manière manuelle. Nous proposons aussi un clustering automatique où deux jonctions sont considérées comme similaires si leur distance d'édition est limitée par certains paramètres.

Le calcul de la distance d'édition se fait en prenant en compte les homopolymères et les décalages, en plus des traditionnelles substitutions, insertions et délétions. Les homopolymères peuvent être dûs à une erreur de séquençage (selon le séquenceur utilisé). Les décalages correspondent à deux $w$-jonctions dont les $w' < w$ premiers nucléotides de la première jonction, sont identiques aux $w'$ derniers nucléotides de la seconde jonction.

Néanmoins cette étape de clustering est laissée à la discrétion de l'utilisateur afin de ne pas biaiser les résultats. Nous faisons l'hypothèse qu'une erreur de séquençage ne pourra pas changer l'ordre de grandeur de la quantification relative d'un clone, car une erreur de séquençage reste un phénomène rare. En revanche, rassembler, par erreur, des jonctions qui diffèrent par exemple d'un homopolymère dans leur zone de N-diversité, peut amener à mettre ensemble des clones qui représentent un fait des lymphocytes différents.

## 3 Résultats et discussion

*Implémentation.* Les algorithmes ont été implémentés dans un logiciel C++ appelé Vidjil. Ce logiciel est mis à disposition, en open-source, à la communauté à l'adresse `http://bioinfo.lifl.fr/vidjil`. Les tests ci-dessous portent sur la version 2013.04 de Vidjil, qui analyse des jeux de données d'un million de reads en moins de 10 minutes sur un ordinateur portable standard. Vidjil est environ deux à cinq fois plus rapide que igBlast et cent fois plus rapide que le serveur web de IMGT/HighV-QUEST, en produisant certes des résultats bien moins détaillés, mais suffisants pour des études de clonalité.

*Jeux de données.* Des échantillons de la moelle osseuse d'un patient ont été séquencés au diagnostic (Diag) et à trois points de suivi (Fu-1, Fu-2 et Fu-4). Afin de tester la robustesse de la méthode, des dilutions du diagnostic ont été effectuées dans des échantillons de cinq personnes saines. Ces dilutions sont appelées en fonction de la concentration du diagnostic Scale-$10^{-2}$, Scale-$10^{-3}$, Scale-$10^{-4}$, Scale-$10^{-5}$. Ces huit échantillons ont été séquencés avec des profondeurs allant de 85 000 à 1 306 606 reads. Le consentement du patient a été recueilli suivant les procédures en vigueur.

| Scale-$10^{-5}$ | Vidjil – igBlast | Vidjil – HighV-QUEST | igBlast – HighV-QUEST |
|---|---|---|---|
| 0 .. 4 | 23562 (90.0%) | 20171 (92.5%) | 21074 (85.7%) |
| 5 .. 9 | 1832 (7.0%) | 1358 (6.2%) | 2865 (11.7%) |
| 10 .. 14 | 512 (2.0%) | 253 (1.2%) | 448 (1.8%) |
| 15 .. 19 | 152 (0.6%) | 17 (0.1%) | 94 (0.4%) |
| $\geq 20$ | 94 (0.4%) | 17 (0.1%) | 94 (0.4%) |

**Table 1.** Comparaison des prédictions des centres des jonctions entre igBlast, IMGT/HighV-QUEST, et l'heuristique de Vidjil, sur les 100 000 premiers reads de l'échantillon Scale-$10^{-5}$ effectivement segmentées. Les valeurs indiquent la distribution de l'écart entre chaque programme. Les prédictions de Vidjil sur les centres des jonctions sont similaires à celles des deux autres logiciels (et d'ailleurs l'écart entre HighV-QUEST et IgBlast est comparable à celui que l'on peut voir avec Vidjil.) Notons que les deux autres outils fournissent beaucoup plus d'information que Vidjil, avec des alignements aux séquences de référence, et, dans le cas de IMGT HighV-QUEST, une analyse détaillée de la jonction.

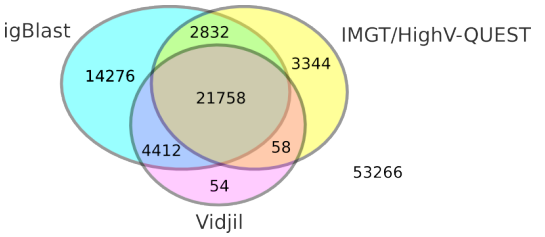



**Figure 3.** Nombre de reads, pour chaque outil, qui ont été effectivement segmentés sur les 100 000 premiers reads de l'échantillon Scale-$10^{-5}$. La quasi totalité des reads segmentés par Vidjil sont également segmentés par un autre outil.

*Evaluation de la prédiction de jonction.* La prédiction de la $w$-jonction repose sur une heuristique qui n'utilise pas la programmation dynamique. Elle peut donc être moins précise qu'un algorithme plus coûteux en temps, qui alignerait la totalité de la séquence. Nous avons donc comparé, sur certains de nos jeux de données, la position du centre de la $w$-jonction donnée par Vidjil, avec les positions déduites des résultats de deux logiciels pris comme référence : IMGT/HighV-QUEST [2] et IgBlast, lancés tous deux sur les séquences de référence de IMGT/GENE-DB [9].

Les résultats (Table 1) montrent que, lorsqu'il y a une prédiction, le centre de la jonction prédite par Vidjil diffère de celui prédit par HighV-QUEST ou igBlast de moins de 10 positions dans plus de 97 % des cas. Elle diffère de moins de 15 positions dans environ 99 % des cas. Avec des jonctions de taille 40, une imprécision de 14 positions n'est pas un problème: il reste suffisament de zone spécifique à chaque réarrangement dans la jonction prédite. Par contre, il serait problématique que la jonction se situe uniquement dans le V ou dans le J, ce qui n'est pas le cas ici. La précision de Vidjil est donc telle que seule une très faible proportion des jonctions pourrait être incorrecte. Cela est confirmé par le fait que seuls 0,2 % des reads segmentés par Vidjil ne le sont pas par un autre outil (voir Fig. 3).

*Abondance des clones.* La Fig. 4 montre la concentration, dans chacun des huit échantillons, de 20 clones (regroupant les 5 clones les plus présents à chaque échantillon). Le clone le plus abondant au diagnostic (#01) était exactement celui connu pour ce patient. Les clones suivants n'avaient pas été suivis par les méthodes conventionnelles. Les clones PBL-1 à PBL-6 sont trouvés uniquement dans les dilutions, à des taux relativement stables, ce qui est un signe de robustesse de la méthode. Comme attendu, on voit dans les diverses dilutions une décroissance du clone #01, mais elle n'est pas calibrée pour les grandes dilutions. Un protocole expérimental amélioré pourrait permettre de mieux comprendre ces biais.

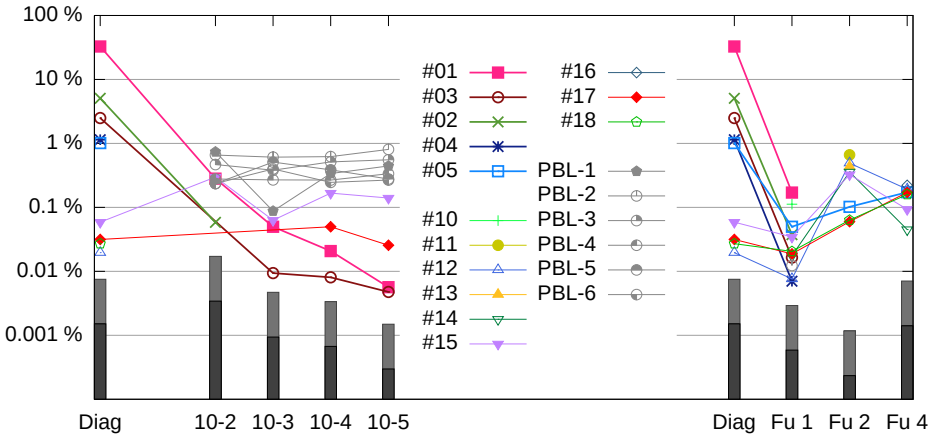



**Figure 4.** Évolution des clones principaux d'un patient, partant du diagnostic et en dilué à concentration décroissantes de $10^{-2}$, $10^{-3}$, $10^{-4}$ et $10^{-5}$ (gauche du graphique) ; et trois points de suivi du patient (Fu1, Fu2, Fu4, partie droite). Les clones #01 à #05 sont les cinq plus abondants détectés au diagnostic, les autres sont parmi les cinq plus abondants dans au moins un échantillon. Les clones PBL-1 à PBL-6 sont trouvés dans au moins deux dilutions, et jamais dans un autre point que la dilution. En bas de chaque graphique, les boîtes noires et grises indiquent la résolution maximale, dépendant du nombre de reads de chaque échantillon (noire: 1 read, gris: 5 reads).

**Conclusion.** La méthode proposée permet d'analyser rapidement des grands jeux de données contenant des réarrangements V(D)J. Appliquée au suivi de leucémie résiduelle, Vidjil suit les variations du clone principal, et identifie d'autres clones qui devraient être analysés, en particulier pour déterminer s'ils sont malins ou non. À terme, il sera possible d'établir une véritable phylogénie entre les sous-clones et ainsi de reconstituer la progression de la maladie ou son évolution après les différentes séquences thérapeutiques.

Les perspectives à long terme dépassent l'étude des syndromes lymphoprolifératifs : quelle est la diversité réelle en termes de clones ou réarrangements de segments V, D et J distincts ? Comment une réaction immunologique se traduit, quantitativement, sur le profil d'un individu ? Au sein d'une population, quelles sont les variations entre les profils de plusieurs individus ?

## Références

[1] IgBlast. `http://www.ncbi.nlm.nih.gov/igblast/`.

[2] Eltaf Alamyar, Véronique Giudicelli, Shuo Li, Patrice Duroux, and Marie-Paule Lefranc. IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and t cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Research*, 8(1), April 2012.

[3] Scott D Boyd, Bruno A Gaëta, Katherine J Jackson, Andrew Z Fire, Eleanor L Marshall, Jason D Merker, Jay M Maniar, Lyndon N Zhang, Bita Sahaf, Carol D Jones, Birgitte B Simen, Bozena Hanczaruk, Khoa D Nguyen, Kari C Nadeau, Michael Egholm, David B Miklos, James L Zehnder, and Andrew M Collins. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol*, 184(12):6986–92, 2010.

[4] Scott D Boyd, Eleanor L Marshall, Jason D Merker, Jay M Maniar, Lyndon N Zhang, Bita Sahaf, Carol D Jones, Birgitte B Simen, Bozena Hanczaruk, Khoa D Nguyen, Kari C Nadeau, Michael Egholm, David B Miklos, James L Zehnder, and Andrew Z Fire. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med*, 1(12):12ra23, 2009.

[5] Xavier Brochet, Marie-Paule Lefranc, and Véronique Giudicelli. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res*, 36(Web Server issue):W503–8, 2008.

[6] JJM Dongen, T Szczepański, and HJ Adriaansen. Immunobiology of leukemia. In Greaves M Henderson ES, Lister TA, editor, *Leukemia, 7th edition*. Saunders, 2002.

[7] Bruno A Gaëta, Harald R Malming, Katherine J L Jackson, Michael E Bain, Patrick Wilson, and Andrew M Collins. iHMMune-align: hidden markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics (Oxford, England)*, 23(13):1580–1587, July 2007. PMID: 17463026.

[8] C. Gawad, F. Pepin, V. Carlton, M. Klinger, A.C. Logan, D.B. Miklos, M. Faham, G. Dahl, and N. Lacayo. Massive evolution of the immunoglobulin heavy chain locus in children with B precursor acute lymphoblastic leukemia. *Blood*, 2012.

[9] Véronique Giudicelli, Denys Chaume, and Marie-Paule Lefranc. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Research*, 33(S1):D256–D261, 2005.

[10] Katherine J L Jackson, Scott Boyd, Bruno A Gaëta, and Andrew M Collins. Benchmarking the performance of human antibody gene alignment utilities using a 454 sequence dataset. *Bioinformatics*, 26(24):3129–30, 2010.

[11] Gunter Kerst, Hermann Kreyenberg, Carmen Roth, Catrin Well, Klaus Dietz, Elaine Coustan-Smith, Dario Campana, Ewa Koscielniak, Charlotte Niemeyer, Paul G. Schlegel, Ingo Müller, Dietrich Niethammer, and Peter Bader. Concurrent detection of minimal residual disease (MRD) in childhood acute lymphoblastic leukaemia by flow cytometry and real-time PCR. *British Journal of Haematology*, 128(6):774–782, 2005.

[12] Aaron C Logan, Hong Gao, Chunlin Wang, Bita Sahaf, Carol D Jones, Eleanor L Marshall, Ismael Buno, Randall Armstrong, Andrew Z Fire, Kenneth I Weinberg, Michael Mindrinos, James L Zehnder, Scott D Boyd, Wenzhong Xiao, Ronald W Davis, and David B Miklos. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc Natl Acad Sci U S A*, 108(52):21194–21199, 2011.

[13] S. Munshaw and T.B. Kepler. SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics*, 26(7):867–872, 2010.

[14] L Ohm-Laursen, M Nielsen, SR Larsen, and T Barington. No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. *Immunology*, 2(119):265–77, 2006.

[15] M.M. Souto-Carneiro, N.S. Longo, D.E. Russ, H. Sun, and P.E. Lipsky. Characterization of the human ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JoinSolver. *The Journal of Immunology*, 172(11):6790, 2004.

[16] J J M van Dongen, A W Langerak, M Brüggemann, P A S Evans, M Hummel, F L Lavender, E Delabesse, F Davi, E Schuuring, R García-Sanz, J H J M van Krieken, J Droese, D González, C Bastard, H E White, M Spaargaren, M González, A Parreira, J L Smith, G J Morgan, M Kneba, and E A Macintyre. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98-3936. *Leukemia*, 17(12):2257–317, 2003.

[17] René L. Warren, J. Douglas Freeman, Thomas Zeng, Gina Choe, Sarah Munro, Richard Moore, John R. Webb, and Robert A. Holt. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Research*, 21(5):790 –797, 2011.

[18] Joshua A Weinstein, Ning Jiang, Richard A White, 3rd, Daniel S Fisher, and Stephen R Quake. High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324(5928):807–10, 2009.

# Complete conservation of human protein tandem repeats across the eukaryotic clade

Elke SCHAPER[1,2], Olivier GASCUEL[2] and Maria ANISIMOVA[1]

[1] D-INFK, D-USYS, ETH Zürich, Universitätstr. 6, 8092 Zürich, Suisse
{Elke, Maria.Anisimova}@inf.ethz.ch
[2] Institut de Biologie Computationnelle, LIRMM, CNRS-UM2, 95 rue de la Galera, 34095 – Montpellier, France
Olivier.Gascuel@lirmm.fr

**Keywords** Tandem repeats, sequence evolution, conservation, comparative analysis.

## 1 Introduction

Tandem repeats (TRs) are consecutive duplicates of genomic sequences. They represent one of the most frequent sequence features in both coding and non-coding DNA. TRs evolve through expansion and deletion of repeat units, with mutation rates found at six orders of magnitude higher than point mutation rates [1]. It has been argued that these high mutation rates may lead to a pool of variation also in protein TRs, constituting a source for rapid adaption to fast changing environments (e. g. [2]). A fast succession of TR unit expansions and deletions on the population scale would eliminate conservation of TRs across species. We examined the conservation of human TRs across the eukaryotic clade, tracing the evolutionary history of every single repeat unit by means of a comparative phylogenetic analysis. Our results reveal a high and complete conservation of a large number of TRs within the eukaryotic clade, thus showing a strong discrepancy with common beliefs on TR evolution.

## 2 Data & Methods

Circular protein HMMs [3,4] were build from PFAM A domains [5] and *de novo* predicted TRs in the human proteome (details in [6]). With these circular HMMs, TR predictions were refined for the human proteome. Overlapping TRs, TRs with units shorter than 15 amino acids, and with fewer than four TR units were discarded.
For the remaining TRs, their corresponding circular HMMs were used to predict homologous TRs on all orthologous genes in 61 Eukaryotic species clustered in Ensembl Compara gene trees [7]. To study the evolutionary history and conservation of TR units, for each orthologous pair of TR-containing genes we reconstructed maximum likelihood phylogenies using PhyML [8] (see FIG. 1 for an example).

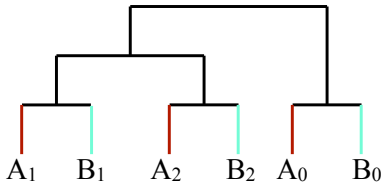



**Figure 1.** Example of TR unit phylogeny for two species A and B with $n_A = n_B = n_c = n_{cb} = 3$ and $k = 1$.

Next, the following measures for conservation of TRs in two orthologous genes were calculated:
- The number of TR units in both genes $n_A$ and $n_B$,
- The number of cherries (pairs of adjacent tips) on the phylogeny $n_c$, and specifically, the number of cherries containing TR units of both species, named bisample cherries, $n_{cb}$.
- The indices of all pairs of tandem repeats within the bisample cherries. E. g., a cherry formed by the first TR unit in one species and the third TR unit in the other species has an index pair of $(1, 3)$. Then, the Kendall rank statistic $k$ on all index pairs was calculated.
- The parsimony score when labelling the tree leaves with species A/B. E. g., for the phylogeny in FIG. 1 we obtain 3, indicating that the speciation A/B followed the duplication of TR in the ancestor of A and B. When the parsimony value is 1, speciation occurred first and was followed by distinct duplications in both species.

Subsequently, we used these measures to establish different degrees of TR conservation, beginning with the most strict definition: As an indicator for *perfect TR unit conservation*, we required that all TR units are found in bisample cherries over the tree and that any $i$th TR unit in one species is paired with the $i$th TR unit in the other species ($n_A = n_B = n_c = n_{cb}$) and $k = 1$. Further, we attribute *strong TR unit conservation*, if ($\max(n_A, n_B) - n_{cb} \leq 1$), $n_A, n_B \geq 4$ and $k = 1$.

## 3 Results

3085 non-overlapping TRs were annotated in 20,163 Ensembl gene families (TRs in 2530 human genes). Of these, 355 TRs are *de novo* annotations. Further, 568 describe Zinc fingers, 225 LRRs, and 165 Ankyrin repeats.

We find that in most cases, the conversation of the gene entails the conservation of several TR units, which indicates relevance for the gene function (FIG. 2: *1-2*). Surprisingly, the majority of human TRs (2780 of 3085, 90.1%) show *perfect TR unit conservation* on the species level (FIG. 2: *4*), suggesting that these TRs have not undergone neither duplications nor deletions at least since the most recent speciation and thus are not subject to neutral expansions or deletions on the population level. This means that advantageous TR unit numbers rapidly spread to fixation, and further TR indel events are inhibited. This strong conservation seems to indicate an important structural and functional role of TRs.

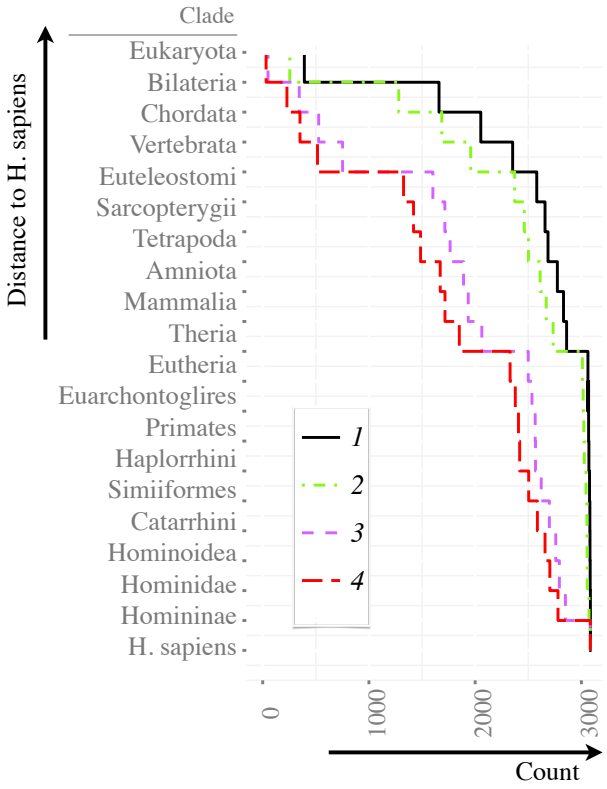



**Figure 2.** Most distantly observed conservation of 3085 human protein TRs across the eukaryotic clade.
*1:* The TR containing gene is present.
*2:* Additionally, $\geq 4$ TR units are present.
*3:* The TR exhibits *strong TR unit conservation*.
*4:* The TR exhibits *perfect TR unit conservation*.

## Acknowledgements

This work was supported by grants from the SNSF (31003A-127325) and the Germaine de Staël program of the Swiss Academy of Engineering Sciences, both to Maria Anisimova.

## References

[1] Ellegren,H. (2000) *Microsatellite mutations in the germline: implications for evolutionary inference.* Trends Genet., 16, 551-558.

[2] Gemayel,R., Vinces,M.D., Legendre,M. and Verstrepen,K.J. (2010) *Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences.* Annu. Rev. Genet., 44, 445-477.

[3] Eddy,S.R. (2000) *Profile hidden Markov models for biological sequence analysis.* Washington University School of Medicine, St Louis, MO (http://hmmer.wustl.edu/).

[4] Uricaru,R., Bréhélin,L. and Rivals,E. (2007) *A new type of Hidden Markov Models to predict complex domain architecture in protein sequences.* JOBIM'07.

[5] Punta,M. et al. (2011) *The Pfam protein families database.* Nucleic Acids Research, 40, D290-D301.

[6] Schaper,E., Kajava,A.V., Hauser,A. and Anisimova,M. (2012) *Repeat or not repeat?–Statistical validation of tandem repeat prediction in genomic sequences.* Nucleic Acids Research, 40.

[7] Flicek,P., et al. (2012) *Ensembl 2012.* Nucleic Acids Research.

[8] Guindon, S.X.P. and Gascuel,O. (2003) *A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood.* Systematic Biology, 52, 696-704.

# PARSEC: a new web platform for the localization and characterization of genomic sites in complete eukaryotic genomes

## PARSEC web platform

Alexis Allot[1], Laetitia Poidevin[1], Raymond Ripp[1], Olivier Poch[1] and Odile Lecompte[1]

[1] Laboratoire de Bioinformatique et Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC) UMR7104 CNRS/INSERM/UDS, 67404, Illkirch, France

{alexis.allot, laetitia.poidevin, raymond.ripp, olivier.poch, odile.lecompte}@igbmc.fr

**Abstract** *We present PARSEC (PAtteRn Search and Contextualization), a new open source platform of guided discovery allowing localization and biological characterization of short genomic sites in complete eukaryotic genomes. PARSEC can search for a genomic sequence or a degenerated pattern with a specified number of mismatches. The set of genomic sites can then be characterized in terms of (i) conservation in model organisms, (ii) genetic context (proximity to genes) and (iii) function of neighboring genes. These modules allow the user to explore, visualize, filter out and extract biological knowledge from a set of short genomic regions such as transcription factor binding sites.*

**Keywords** Genomic site, pattern search, biological context, webserver

## PARSEC: une nouvelle plateforme web pour la localisation et la caractérisation de sites génomiques dans les génomes eucaryotes complets.

### PARSEC plateforme web

**Résumé** *Nous présentons PARSEC, une nouvelle plateforme modulaire open source de découverte guidée, permettant la localisation et la caractérisation de sites génomiques courts dans des génomes eucaryotes complets. PARSEC peut rechercher une séquence génomique ou un pattern dégénéré avec un nombre de mismatches spécifié. L'ensemble de sites génomiques peut ensuite être caractérisé d'un point de vue évolutif (conservation dans des organismes modèles), de proximité génétique et de fonction des gènes proches. Ces modules permettent à l'utilisateur d'explorer, visualiser, filtrer et extraire la connaissance biologique d'un ensemble de sites génomiques, comme par exemple les sites de fixation des facteurs de transcription.*

**Mots-clés** Site génomique, recherche de pattern, contexte biologique, serveur web

## 1 Introduction

Short nucleotidic sequences called signals or sites playing crucial role in dynamic gene expression, splicing, replication, DNA repair… [1] are generally located in genomic regions closely related to genes but can also be found in very distal regions [2]. Thousands of these functional sites have been identified in the genomes of various animal models and recent genome-wide association studies in humans have identified thousands of such sites that are strongly associated with a complex disease or a related trait [1]. Thus, identification and annotation of these sites are crucial not only for the understanding of major biological processes but also for human disease analysis and therapies.

By providing huge amounts of genomic and transcriptomic data, the Next Generation Sequencing technologies are revolutionizing the study of genomics sites allowing more statistically sounded localization and characterization of these signals which can be conserved in a given species or between species.

Classically, these sites are represented either as a base frequency matrix or as a Hidden Markov Model (HMM) with context sensitive position frequencies. These statistical representations can only be robustly defined from very large sets of experimentally established sites and can not specify strict conservation of some positions based on biological evidence. An alternate representation is the degenerated pattern, which does not require precision about base frequencies and allows simplistic description of strictly conserved, ambiguous or variable positions. This type of pattern is encoded using IUPAC (International Union of Pure and Applied Chemistry) nomenclature. Pattern searches can easily be performed at complete genome level but generate numerous false positives, thus implying that additional knowledge-based analyses are required to strengthen the final results.

In this context, we have designed a web service, PARSEC, for the rapid IUPAC oriented localization of genomic sites coupled to their characterization at genomic, phylogenetic and functional annotation levels. PARSEC has a modular architecture to propose both search and contextualization aspects in a unique and intuitive interface, guiding the user in discovery paths. The modules propose various analysis steps to explore, visualize, filter out the noise and extract the biological knowledge from large sets of sites.

The program exploits an efficient data structure, the compressed suffix trees (CSTs) [3], to rapidly localize degenerated patterns in complete eukaryotic genomes (8 genomes available for now: human, mouse, rat, chicken, zebrafish, drosophila, nematode, yeast). This search module is coupled to four characterization modules using:

- The phylogenetic conservation for selective filtering of sites conserved in target genomes with three levels of stringency.
- The genetic context analysis, for selection of sites according to their distances to neighboring genes and according to the neighboring gene types (tRNA, snRNA, snoRNA, miRNA, protein coding…).
- The functional enrichment based on GoMiner [4] for the characterization of genomic sites in terms of biological process, molecular function and cellular component.
- The functional filtering for retrieval of sites nearby genes of a given biological function.

Importantly, instead of proposing a catalogue of tools, we offer at each step of the analysis only the tools relevant to previously produced results, combining modules in analysis protocols adapted to precise biological questions.

## 2 Concept of PARSEC

PARSEC approach is designed for i) degenerate pattern oriented search, ii) the use of local data for performance and robustness purposes, iii) the possibility to apply scenarii devoted to specific biological questions, iv) a modular and extensible structure with nonlinear execution, v) open source code, and vi) the possibility to query through web service or to download and install it on a local server.

This approach relies on two main systems: a Data Manager to import, integrate and update huge amounts of biological data, and a Web Service allowing the user to run analyses and view the results.

### 2.1 Managing the genomic information database with Parsec Data Manager

The DataManager relies on Apache Commons Net to retrieve files from distant FTP servers, and on Commons Compress to deal with specific files. Alignments and genes are stored in a Postgres database.

- **Genomic sequences:** FASTA sequences of complete masked genomes are retrieved from the UCSC [5]. For the genomic sites search, the use of masked genomes allows to filter out numerous potential false positive hits from repeated regions, at the expense of some true positives. Using the compressed suffix trees

library [6], one CST is built for each chromosome and saved. The CST is saved into three files, '.bp' (balanced parenthesis, representing the structure of the tree), '.csa' (the compressed suffix array) and '.lcp'(the longest common prefix of suffixes). An additional file with sequence length is used for loading of the CST in main memory.

- **Alignments:** Evolutionary conservation characterization is estimated using the pairwise alignments produced by blastZ [7] and retrieved from the UCSC [5]. Retrieved alignments between human and 48 organisms represent for example more than 53 million table entries.

- **Genes:** Genetic information is retrieved from the EnsGene table from UCSC taking advantage of the high number of organisms available and of the relative completeness in terms of gene types (protein coding, rRNA, tRNA, snRNA, miRNA, snoRNA, etc.). Extracted data is updated with transcript type information from EnsemblSource table and official gene name information from ensemblToGeneName table. Because tRNA data lack for several organisms, ensGene data is complemented for human, mouse, rat and zebrafish with information from tRNAs table.

- **Gene Ontology:** We used a script available with High Throughput GoMiner package [4] to install the most up to date GO database. These tables are used by GoMiner for functional enrichment.

## 2.2 Localization and characterization of genomic sites with PARSEC Web Service

PARSEC is based on servlet technology and runs on an Apache Tomcat server. PARSEC includes 5 modules:

- **Search**:

The compressed suffix trees approach has been preferred as a data structure. CSTs exhibit properties particularly suited for genomic data, such as quick retrieval of sequence corresponding to chromosomal coordinates, search for the most repeated sequence of a given length, search for the longest common sequence of given chromosomes or genomes, etc. Among the distinct CST implementations, we used the one developed by the laboratory of succinct data structures (SUDS) [6] since it is programmed in C++ and offers simple methods for manipulating and navigating through the simulated suffix trees. We have extended the existing library by implementing recursive tree navigation for fast and easy degenerated pattern search. The degenerate nature of IUPAC pattern is particularly adapted for tree structure navigation, each ambiguity symbol resulting in the traversal of several edges of the tree at the same time. For example, when searching the pattern ATRA corresponding to AT[AG]A, the beginning of the pattern (AT) will be considered only once, then the navigation function will progress at the same time on the edge corresponding to A and the edge corresponding to G. This avoids exact mapping of all possible combinations corresponding to a degenerated pattern.

In addition, we implemented chromosomal search parallelization (managing an array of CST representing chromosomes) on the JAVA side.

To ensure equitable use of the service, submitted patterns must at least be 5 base pairs long and have a reasonable complexity.

- **Evolutionary conservation**:

The evolutionary conservation module can take as input a set of sites queried in a specified genome (query sites) and provided by compatibles modules of PARSEC (pattern search, genetic context, evolutionary conservation, functional filtering, see below). The user can retain limited set of query sites by defining a set of target organisms coupled to evolutionary conservation criteria. He can then select the minimum number of organisms in which a site must be conserved to be kept. PARSEC allows 3 hierarchized levels of conservation stringency: (I) ALIGNED_SITE: The query site is found in a region conserved between the query and the target genomes (i.e. region aligned in the BLASTZ pairwise alignments available at the UCSC) and is perfectly aligned with a target site; (II) CONSERVED_SITE: a query site is found in a conserved target region exhibiting a target site, but alignment of the query and target sites is not required; (III) ALIGNED_REGION: The query site is located in a region conserved in the target genome (Fig 1).

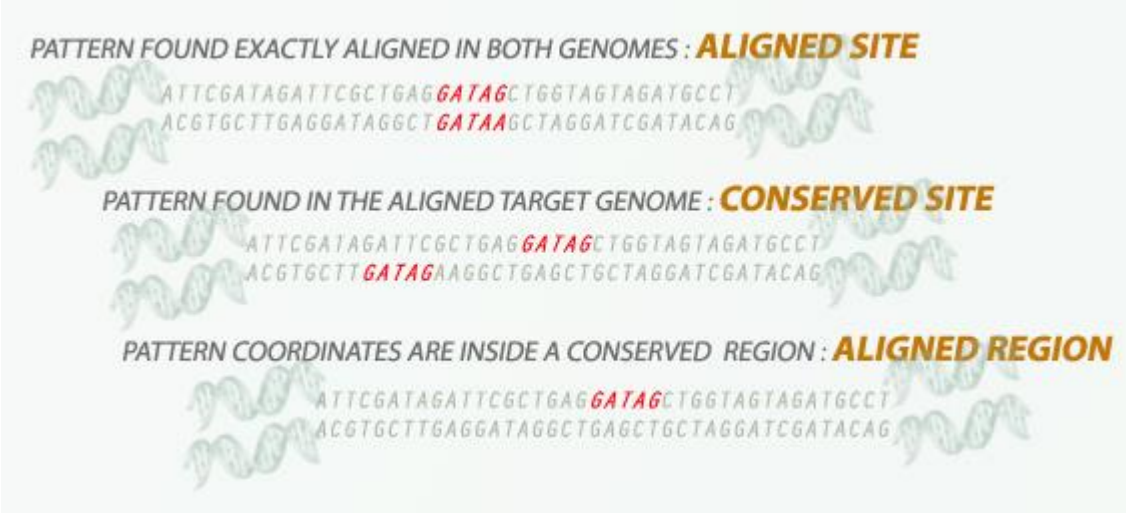
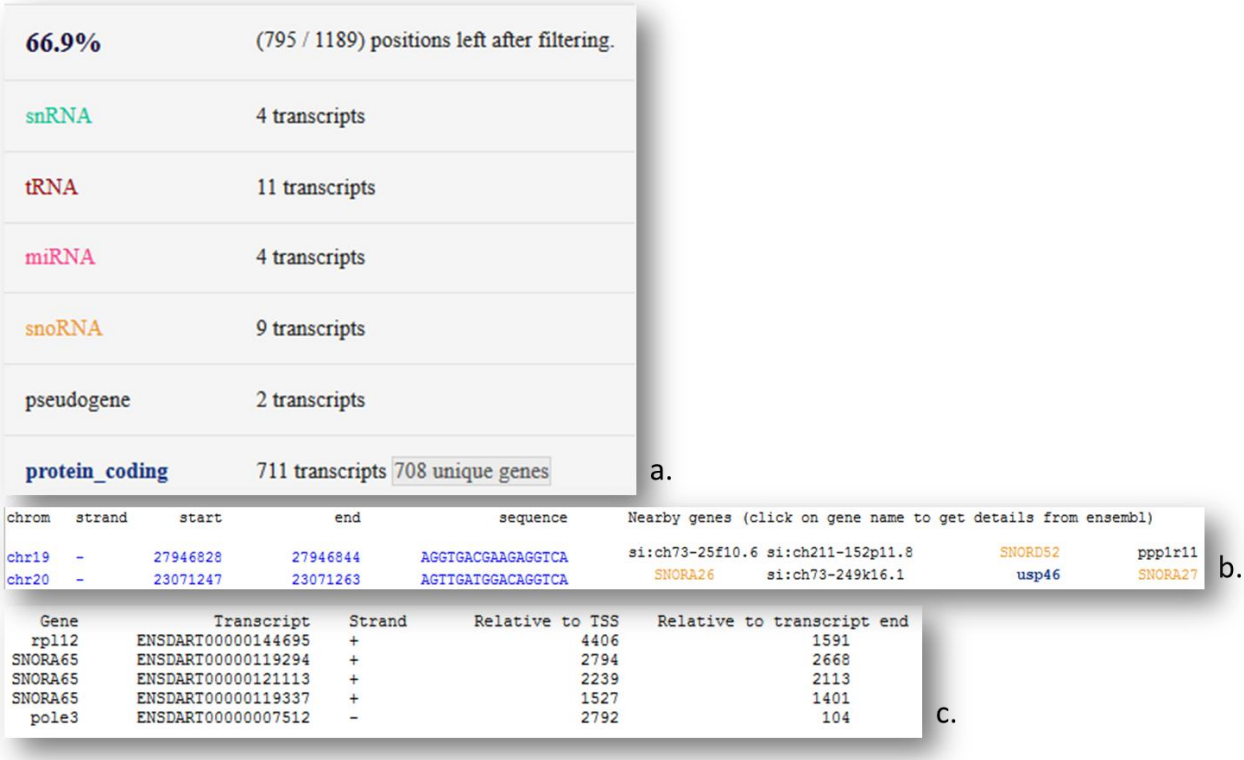



**Figure 1.** PARSEC allows 3 hierarchized levels of conservation stringency (see main text). In this example the searched pattern is GATAR (R meaning A or G)

- **Genetic proximity:**

The genetic module allows the detection of gene candidates spatially linked to the genomic site of interest with a high level of customization. Selection and subsequent analysis can be done relatively to the two gene boundaries or to the Transcription Start Site (TSS). The user can select the sites close to specific gene types including miRNA, rRNA, tRNA, protein-coding gene. An overview of the distinct gene types (Fig. 2.a.) is provided. The user can see the gene(s) detected near each site (Fig 2.b.) and distances between site and each gene (including variants) (Fig. 2.c.). In addition, links to the UCSC genome browser offer further graphical exploration.



| | |
|---|---|
| **66.9%** | (795 / 1189) positions left after filtering. |
| snRNA | 4 transcripts |
| tRNA | 11 transcripts |
| miRNA | 4 transcripts |
| snoRNA | 9 transcripts |
| pseudogene | 2 transcripts |
| **protein_coding** | 711 transcripts 708 unique genes |

a.

| chrom | strand | start | end | sequence | Nearby genes (click on gene name to get details from ensembl) | | | |
|---|---|---|---|---|---|---|---|---|
| chr19 | - | 27946828 | 27946844 | AGGTGACGAAGAGGTCA | si:ch73-25f10.6 | si:ch211-152p11.8 | SNORD52 | ppp1r11 |
| chr20 | - | 23071247 | 23071263 | AGTTGATGGACAGGTCA | SNORA26 | si:ch73-249k16.1 | usp46 | SNORA27 |

b.

| Gene | Transcript | Strand | Relative to TSS | Relative to transcript end |
|---|---|---|---|---|
| rpl12 | ENSDART00000144695 | + | 4406 | 1591 |
| SNORA65 | ENSDART00000119294 | + | 2794 | 2668 |
| SNORA65 | ENSDART00000121113 | + | 2239 | 2113 |
| SNORA65 | ENSDART00000119337 | + | 1527 | 1401 |
| pole3 | ENSDART00000007512 | - | 2792 | 104 |

c.

**Figure 2.** Genetic module allows the user to overview the population of detected genes and transcripts (a), the genes detected near a site (b) and the distances between a selected site and detected nearby transcripts (c).

- **Functional Enrichment** :

For the functional enrichment module, we have created a layer allowing the use of GoMiner as a simple java library. It allows the functional characterization of previously identified set of genes close or encompassing query sites. This can be for example useful for identification of process regulated by a transcription factor.

- **Functional Filtering** :

The functional filtering step allows the selection of sites probably related to specific molecular functions, biological process or cellular components.

## 3 Using PARSEC

PARSEC can be used for various exploitation scenarios. For instance, the user can retrieve the set of genes potentially regulated by a transcription factor (Fig. 3). We searched for the DR5 (Direct repeats separated by 5 bp) Retinoic Acid Response Element (RARE) in the masked human genozme (version hg19), using the consensus pattern RGKTSANNNNNRGKTSA [8]. We localized 14,251 sites in 5.3s. Using the conservation module, we then selected the sites conserved ("Aligned site" and "Conserved site" parameters) in at least one of the following vertebrate species: *Xenopus tropicalis, Danio rerio, Takifugu rubripes, Tetraodon nigroviridis, Oryzias latipes, Gasterosteus aculeatus, Petromyzon marinus*. We found 178 sites (execution time of 20.8s). Among them, we filtered the 104 sites located near a Transcription Start Site (5000 bp upstream, 5000 bp downstream) in the human genome using the nearby genes module (execution time: 0.13 s). We then analyzed the nearby protein coding genes with the Functional Enrichment module. This identified several GO categories with a significant enrichment (False Discovery Rate<10-5) among which: embryo development, embryonic organ development and retinoic acid receptor signaling pathway. These categories are known to be related to retinoic acid regulation [9], which demonstrates PARSEC ability to produce biologically meaningful genomic site analysis.

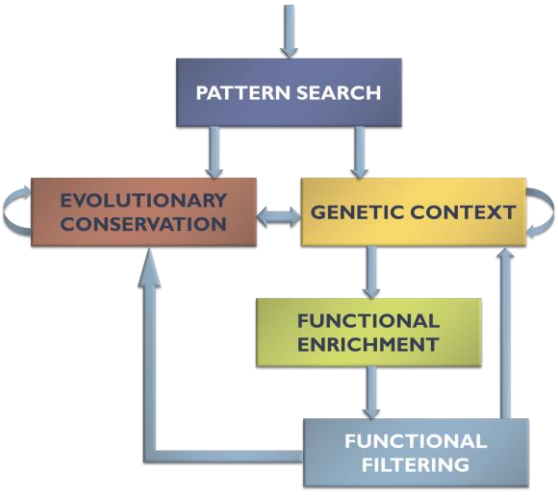



**Figure 3.** Modules and their connections are presented. The scenario to define the set of target genes of a transcription factor requires successive execution of PATTERN_SEARCH, EVOLUTIONARY_CONSERVATION, GENETIC_CONTEXT and FUNCTIONAL_ENRICHMENT modules.

## 4 Comparison between PARSEC and other web services

PARSEC has been compared to some web services allowing the localization of DNA sequences in genomes and/or the characterization of the genomic site occurrences.

- BLAT [10] allows fast search in numerous genome sequences but is limited to non-degenerated query sequences with a minimal length of 25bp.

- TagScan [11] provides fast degenerated pattern search (IUPAC) in fourteen genomes, but the display of the results is not adapted for a big number of hits and no additional characterization is done.

- Great [12] offers near genes and enrichment analysis with great statistics in three genomes but does not allow pattern search or conservation characterization.

- GOMO(MEME) [13] identifies biological functions of DNA motifs in five genomes, but works only on frequency matrixes, has a limited conservation filtering, and considers only upstream gene regions.

PARSEC stands out by its integrative approach, in an all-in-one modular analysis platform.

## 5 Conclusion

We designed PARSEC to answer a complete genome analysis problem, be modular and easily extensible, and present a biologist-friendly web interface. To improve its analysis abilities and scope, it will be enhanced in several ways.

Frequency matrix and logo will be generated from the set of filtered-out sites. Results produced by the genetic module will be statistically characterized in terms of hits positions relative to gene boundaries, allowing further characterization of selected intervals of sites.

New modules will be added to allow new levels of contextualization for genomic sites. In particular, we will add data about polymorphism (SNP, associated diseases...) and epigenomics (filtering of accessible sites with given tissue/time parameters).

## Acknowledgements

The authors are grateful to Vincent Laudet and Cécile Rochette-Egly for helpful discussions throughout the development of PARSEC.

Funding: This work was supported by the Agence Nationale de la Recherche (grants Puzzle-Fit: 09-PIRI-0018-02 and BIP:BIP: ANR10-BINF03-05).

## References

[1] T. S. Furey, ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions, *Nature reviews. Genetics*, vol. 13, pp. 840-52, Dec 2012.

[2] G. E. Zentner and P. C. Scacheri, The chromatin fingerprint of gene enhancer elements, *The Journal of biological chemistry*, vol. 287, pp. 30888-96, Sep 7 2012.

[3] K. Sadakane, Compressed Suffix Trees with Full Functionality, *Theory of Computing Systems*, vol. 41, pp. 589-607, 2007.

[4] B. R. Zeeberg, H. Qin, S. Narasimhan, M. Sunshine, H. Cao, D. W. Kane, M. Reimers, R. M. Stephens, D. Bryant, S. K. Burt, E. Elnekave, D. M. Hari, T. A. Wynn, C. Cunningham-Rundles, D. M. Stewart, D. Nelson, and J. N. Weinstein, High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID), *BMC bioinformatics*, vol. 6, p. 168, 2005.

[5] L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent, The UCSC Genome Browser database: extensions and updates 2013, *Nucleic acids research*, Nov 15 2012.

[6] N. Valimaki, W. Gerlach, K. Dixit, and V. Makinen, Compressed suffix tree--a basis for genome-scale sequence analysis, *Bioinformatics,* vol. 23, pp. 629-30, Mar 1 2007.

[7] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller, Human-mouse alignments with BLASTZ, *Genome research,* vol. 13, pp. 103-7, Jan 2003.

[8] S. Lalevee, Y. N. Anno, A. Chatagnon, E. Samarut, O. Poch, V. Laudet, G. Benoit, O. Lecompte, and C. Rochette-Egly, Genome-wide in silico identification of new conserved and functional retinoic acid receptor response elements (direct repeats separated by 5 bp), *The Journal of biological chemistry,* vol. 286, pp. 33322-34, Sep 23 2011.

[9] S. Kumar and G. Duester, Retinoic acid signaling in perioptic mesenchyme represses Wnt signaling via induction of Pitx2 and Dkk2. *Developmental biology,* vol. 340, pp. 67-74, Apr 1 2010.

[10] W. J. Kent, BLAT--the BLAST-like alignment tool, *Genome research,* vol. 12, pp. 656-64, Apr 2002.

[11] C. Iseli, G. Ambrosini, P. Bucher, and C. V. Jongeneel, Indexing strategies for rapid searches of short words in genome sequences, *PloS one,* vol. 2, p. e579, 2007.

[12] C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano, GREAT improves functional interpretation of cis-regulatory regions, *Nature biotechnology,* vol. 28, pp. 495-501, May 2010.

[13] F. A. Buske, M. Boden, D. C. Bauer, and T. L. Bailey, Assigning roles to DNA regulatory motifs using comparative genomics, *Bioinformatics,* vol. 26, pp. 860-6, Apr 1 2010.

# Session 6B : Évolution

# Evolutionary dynamics of *Escherichia/Shigella* Fimbriome

Virginie CALDERON[1], Yves QUENTIN[1], Sophie de BENTZMANN[2]and Gwennaele FICHANT[1]

[1] Laboratoire de Microbiologie et Génétique Moléculaires, UMR5100 CNRS/UPS, 118 route de Narbonne, 31062, Toulouse, Cedex 9, France

{calderon, quentin, fichant}@ibcg.biotoul.fr

[2]Laboratoire d'Ingénierie des Systèmes Macromoléculaires, UMR7255 CNRS/Aix-Marseille Université, 31chemin Joseph Aiguier, 13402, Marseille, France

bentzmann@imm.cnrs.fr

***Abstract*** *Bacterial interactions with environmental surfaces or host tissues depend of their ability to adhere to target surface. This is mediated through various surface-exposed adhesive systems. Among those systems, the Chaperone-Usher (CU) pathway forms a large multigenic type 1 fimbriae family that can be classify into six major subfamilies. Previous analyses of complete genomes from different strains of Escherichia coli/Shigella had revealed large variability in CU system content. Here, we address the evolutionary dynamics of Escherichia/Shigella fimbriome in 40 genomes from E. coli and Shigella. The distribution of the CU systems is not correlated to the strain classification into pathovars. Most of the systems are predicted to be acquired only once along the history of Escherichia/Shigella strains. After their insertion, most of them have suffered different mutational events (pseudogenisation and recombination) leading to their loss of function. This is especially at work in Shigella strains were only two systems are kept functional. Evolutionary dynamics of these systems in Escherichia may reflect a rapid strain adaptation to new environments. The most probable scenarios suggest a site-specific integration or excision of these systems through a molecular mechanism that remains to be elucidated.*

**Keywords** Evolution, Dynamic, Fimbriome.

## 1 Introduction

L'interaction des bactéries avec les différentes surfaces de leur environnement et les tissus de l'hôte dépend de leur capacité à adhérer à une surface cible. Cette étape essentielle pour la colonisation et l'infection par des bactéries pathogènes est médiée par l'exposition à la surface des bactéries de systèmes d'adhésion. Chez les bactéries à Gram-négatif, l'arsenal de systèmes d'adhésion est composé de systèmes de sécrétion de type 1 (SSTI), d'autotransporteurs (SSTV), de pili (SST IV), de curli, de flagelle et de fimbriae assemblé par la voie chaperonne-usher (CU). Les systèmes assemblés par la voie CU forment une large famille multigénique de fimbriae.

Un système CU est constitué d'au moins trois types de protéines différents, une protéine de la membrane externe appelée usher qui forme le pore de translocation, une chaperonne périplasmique et une sous-unité structurelle majeure de la fimbriae. Cependant, la majorité des systèmes présente d'autres protéines, comme les sous-unités mineures, les adhésines ou des chaperonnes supplémentaires. Les gènes codant pour les partenaires des systèmes CU sont généralement organisés en opéron. Une classification non-ambigüe des systèmes CU a été proposée [1] sur la base de l'analyse phylogénique des protéines usher, seul composant unique de ces systèmes dont la séquence apparaît conservée au travers de cette famille. Six grandes sous-familles ont été identifiées, appelées α-, β-, γ-, κ-, π- et σ-fimbriae. La sous-famille γ a été décomposée en quatre sous-sous-familles ($\gamma_1$-, $\gamma_2$-, $\gamma_3$- et $\gamma_4$-). Ces systèmes peuvent être portés par des plasmides ou codés sur les chromosomes, parfois associés à des ilots de pathogénie (PAI).

Parmi les bactéries à Gram-négatif, les Enterobacteriaceae codent pour un grand nombre de systèmes CU, et plus particulièrement l'espèce *Escherichia coli* [2,3] chez qui la distribution de ces systèmes varie entre les différentes souches. Les génomes complets d'un grand nombre de souches d'*Escherichia/Shigella* étant disponibles, nous avons entrepris l'étude des différentes dynamiques d'évolution des systèmes CU au sein de ces génomes. Un travail publié alors que ce projet était déjà bien avancé propose un inventaire de ces systèmes dans 36 souches de *E. coli*, leur classification en sous-familles ainsi que la distinction entre

systèmes « intact » et « interrompu » (présence d'une IS ou de pseudogènes dans l'opéron) [2]. Notre analyse a conduit à l'identification des mêmes groupes de systèmes que ceux publiés. Cependant, notre démarche d'annotation étant plus détaillée, nous avons pu identifier précisément les évènements mutationnels qui ont affecté chaque système (nombre de pseudogènes, gènes interrompus, insertion d'IS) et aborder la dynamique évolutive de ces systèmes. L'alignement réalisé des 40 génomes des souches d'*Escherichia/Shigella* nous a permis d'analyser le contenu du locus d'insertion de chaque groupe de systèmes et de mettre en évidence l'existence de systèmes CU présents sous forme de vestiges, *i.e.*, seuls des petits fragments du système sont identifiables ainsi que la présence d'autres gènes au locus. Cette annotation précise a rendu possible l'étude de la dynamique évolutive de ces systèmes dans les souches d'*Escherichia/Shigella*, par l'analyse du flux de gènes et par la détection des évènements de recombinaison qui ont pu affecter le système après son acquisition.

## 2 Matériel et Méthodes

Les 60 génomes complets utilisés dans cette étude ont été téléchargés à partir du site EBI (http://www.ebi.ac.uk/genomes). Ils comprennent les génomes complets de 39 *E. coli/Shigella* et 21 génomes de bactéries proches (18 génomes de *Salmonella*, deux de *Citrobacter* et un de *E. fergusonii*). Nous avons utilisé RPS-Blast pour annoter les protéines de chaque génome en utilisant les profils COG et Pfam disponibles dans la base de données CDD du serveur NCBI que nous avons installés en local.

### 1.1 Annotation des Systèmes CU

Pour reconstruire les systèmes CU, nous avons retenu les protéines de nos génomes d'intérêt présentant le meilleur score d'alignement par RPS-Blast avec un profil COG et/ou Pfam correspondant aux domaines caractéristiques d'un des composants des systèmes CU, à savoir : COG3188 pour le usher ; COG3539 et Pfam7434 (sous-famille α) pour l'adhésine ; COG3539, Pfam06551 (sous-famille β) et Pfam0449 (sous-famille α) pour les pilines ; et le COG3121 pour les chaperonnes. Les systèmes CU sont ensuite reconstruits en repositionnant les partenaires du système sur le génome.

Pour compléter cette annotation dans le cas où le gène codant un des partenaires du système était annoté comme pseudogène ou n'avait pas été annoté, nous avons choisi, pour chaque groupe de systèmes, un système de référence dont la séquence ADN a été utilisée pour une recherche par similarité sur les 60 génomes sélectionnés avec le programme BlastN de la suite Blast [3].

### 1.2 Construction d'un Arbre Phylogénétique des protéines Usher

L'alignement des 635 séquences protéiques du usher est réalisé en utilisant MUSCLE [6]. L'arbre est calculé par une méthode de maximum de vraisemblance PhyML [5] avec le modèle de substitution des acides aminés LG sélectionné par ProtTest [8]. La proportion de sites invariants, la fréquence des acides aminés et la distribution gamma sont estimées à partir des données. Nous avons gardé le meilleur des arbres calculé par les algorithmes NNI (Nearest Neighbor Interchange) et SPR (Subtree Pruning and Regrafting). La topologie de l'arbre et la longueur des branches ont été optimisées. Nous avons réalisé 100 bootstrap non paramétriques.

### 1.3 Arbre des Espèces Escherichia-Shigella-Citrobacter-Salmonella

Nous avons sélectionné les familles de COG représentées par une séquence unique dans chacun des 60 génomes et s'alignant avec au moins 80% du profil. Comme critère additionnel, nous avons imposé que toutes les paires de séquences soient des orthologues 1:1. Un ensemble de 549 groupes d'orthologues a ainsi été obtenu. Pour chaque groupe, les séquences ont été alignées avec MUSCLE [6]. Les alignements obtenus ont été nettoyés avec trimAl [7] en utilisant la méthode *automatique1* avec comme critères pour conserver l'alignement: i) la moyenne d'identité de séquence par colonne aligné >0,8 et, ii) la fréquence de colonne dans l'alignement sans gap >80. 517 familles de COG ont été retenues et les alignements concaténés (171768 sites). L'arbre phylogénétique a été calculé avec PhyML [5] en utilisant le modèle d'évolution des protéines JTT+G sélectionné par ProtTest [8]. La proportion de sites invariants et les paramètres de la distribution gamma ont été estimés à partir des données. Le nombre de catégories de taux de substitution a été fixé à huit. La topologie de l'arbre et la longueur des branches ont été optimisées et la topologie améliorée avec l'algorithme NNI. Nous avons réalisé 100 bootstrap non paramétriques.

### 1.4 Détection des Evènements de Recombinaison Non-Homologue

Le logiciel Alien Hunter (AH) a été utilisé (paramètres par défaut) [9] pour identifier, dans les 40 génomes d'*Escherichia/Shigella,* les régions qui ont pu être acquises par transfert horizontal. Ces régions présentent une composition en séquence inusuelles en terme de *k*-mers pour plusieurs valeurs de *k*. Une région sera prédite d'origine étrangère si le score calculé par AH est supérieur à une valeur seuil dépendante du génome et automatiquement calculée par la méthode en fonction de la composition en *k*-mers du génome. Pour compenser la faible spécificité reconnue de AH [10], nous avons recherché, aux extrémités de la région prédite étrangère, la présence d'un gène codant soit pour un ARNt, soit pour une intégrase.

### 1.5 Scenario d'Acquisition des Systèmes CU

Les scénarios les plus probables de gain et de perte des systèmes CU ont été estimés à l'aide d'une méthode de reconstruction des états ancestraux présents aux nœuds de notre arbre de référence des espèces. La méthode de vraisemblance implémentée dans Mesquite 2.75 [11] a été utilisée. Comme nous ne savons pas, *a priori*, si les gains et pertes de gènes se produisent à la même fréquence, nous avons appliqué le modèle d'évolution asymétrique qui permet d'estimer ces deux taux à partir des données. Nous avons construit une matrice dont les lignes correspondent aux souches et les colonnes aux systèmes en décrivant deux états de caractères : 1 pour la présence du système CU et 0 pour son absence. Un système a été considéré comme présent dès lors que des vestiges du système ont pu être identifiés. Un système a été noté absent si aucune trace du système n'est détectée au locus ou si d'autres gènes y sont présents. Les systèmes CU codés sur des îlots de pathogénie ne suivant pas un héritage vertical ont été écartés de cette analyse.

### 1.6 Détection des Evènements de Recombinaison Homologue

La détection des évènements de recombinaison homologue a été réalisée à l'aide de ClonalOrigin [12]. Comme les gènes codant les différents groupes de systèmes CU ne sont pas présents dans l'ensemble des 40 génomes d'*Escherichia/Shigella*, nous avons procédé en deux étapes. Dans un premier temps, nous avons utilisé les régions cœurs alignées de nos 40 génomes pour estimer avec ClonalOrigin la valeurs des trois paramètres suivants : la longueur moyenne de l'événement de recombinaison $\delta$, le taux de mutation $\theta$ et le taux de recombinaison $\rho$. Dans un second temps, ClonalOrigin a été appliqué sur l'alignement de chaque groupe de systèmes CU en utilisant les valeurs des paramètres précédemment calculées.

En pratique, ClonalOrigin n'acceptant pas en entrée un arbre des espèces, la généalogie des 40 souches étudiées est calculée en appliquant ClonalFrame 1.2 [13] sur l'alignement des génomes réalisé avec progressiveMauve (version Mauve 2.3.1) [14]. Nous avons vérifié que cette généalogie était congruente avec notre arbre de référence des espèces. Les 911 blocs d'alignement correspondant aux régions cœurs d'une longueur d'au moins 500 pb ont été extraites des résultats de progressiveMauve. Une première analyse avec ClonalOrigin a été réalisée indépendamment sur chacune de ces régions permettant de calculer, pour chaque région, les valeurs des trois paramètres. Les valeurs obtenues des médianes ($\delta$=317pb, $\theta$=0,0223 et $\rho$=0,0216) sont du même ordre de grandeur que celles calculées dans une étude précédente sur 27 génomes d'*E. coli* [15]. Dans un deuxième temps, pour identifier les évènements de recombinaison homologue pouvant affecter les systèmes CU, ClonalOrigin a été appliqué de façon indépendante sur l'alignement des régions contenant les systèmes CU. Pour conforter les résultats, ClonalOrigin a été exécuté cinq fois de façon indépendante. Chaque exécution consiste en 3 000 000 d'itérations dont 1 000 000 sont éliminées comme "burn-in". Les résultats des cinq exécutions ont été compilés. Une position de l'alignement peut être ou non incluse dans un évènement de recombinaison et l'origine de cet évènement dans la généalogie peut être identique ou différente. Pour chaque position de l'alignement, nous avons regroupé les évènements de recombinaison détectés au cours des différentes itérations, quelque soit leur origine. La fréquence des évènements de recombinaison à une position donnée est ce nombre divisé par le nombre d'itérations. Nous avons sélectionné les évènements de recombinaison fréquents comme étant les régions pour lesquelles toutes les positions ont une fréquence supérieure à 50% sur au moins 300 nucléotides, valeur proche de celle estimée pour $\delta$, avec au moins une position présentant une fréquence supérieure à 70%.

# 3 Fimbriome de *Escherichia/Shigella*

## 3.1 Génomes Analysés

Les 39 génomes étudiés de *E. coli/Shigella* contiennent des souches commensales ou pathogènes correspondant aux différents pathotypes regroupés en deux grandes classes, les souches de *E. coli* pathogènes intestinales (InPEC) et les souches pathogènes extraintestinales (ExPEC) [16]. Cet ensemble de souches englobe également les différents phylogroupes bien connus de *E. coli/Shigella* (A, B1, B2, D, E, S, S1, S2 et S3) [17]. Pour inférer l'acquisition ou la perte d'un système, nous avons ajouté 21 génomes de bactéries proches de *E. coli/Shigella* qui ont servi de groupes externes : un génome de *E. fergusonii*, deux génomes de *Citrobacter* et dix-huit génomes de *Salmonella*.

## 3.2 Annotation et Classification des Systèmes CU

Les locus CU ont été classés en sept catégories (Table 1): i) 345 systèmes « complets » lorsque tous les gènes du systèmes sont intacts, ii) 83 systèmes « altérés » dont un seul des gènes est endommagé et qui pourrait être toujours fonctionnel par complémentation avec un autre système [18], iii) 99 systèmes « très altérés » et a priori non-fonctionnels, iv) 35 « vestiges » de systèmes, v) « absent » lorsqu'aucune trace de gène n'est trouvée entre les bornes flanquantes conservées, vi) 14 possédant uniquement des « IS » au locus, et vii) 31 locus codant pour des « autres gènes » au locus. Les souches de *Shigella* qui sont des pathogènes intracellulaires de primates se distinguent par un nombre très faible de systèmes complets (2 systèmes). Ce résultat est en accord avec des observations précédentes et serait probablement dû au processus d'évolution réducteur touchant ces génomes.

La classification des séquences usher dans les génomes de *E. coli*/*Shigella* en 5 sous-familles (α, β, π, γ et κ) a été réalisée par la reconstruction d'un arbre calculé sur l'alignement multiple des séquences issues de notre échantillon auxquelles nous avons ajouté les séquences utilisées dans l'article de Nuccio et Baumler [19]. A l'aide de cet arbre, nous avons défini 21 groupes de systèmes orthologues, trois systèmes avec un unique représentant (*pix*, *γ2-like* et *K99*) et un groupe incluant des gènes paralogues (*pap*). L'analyse de l'organisation génétique des systèmes permet d'identifier que pour chaque groupe, l'organisation est strictement conservée sauf pour *lpf2* où une seconde adhésine est parfois présente (phylogroupe E) (Table 1). Les gènes codant pour les adhésines (a) sont généralement localisés en 3' de l'opéron, alors que les gènes codant pour les chaperonnes (c) sont majoritairement localisés en 5' de l'opéron. Ces deux gènes encadrent le gène codant pour le usher (u), excepté dans les systèmes π. Cette organisation peut être une réminiscence de l'organisation du système ancêtre ou refléter des contraintes comme celles liées à la stœchiométrie des partenaires dans le système qui pourrait être contrôlé aux niveaux transcriptionnel/traductionnel. Nous pouvons observer que pour chaque famille, les opérons codant les systèmes sont généralement complets ou peu altérés sauf les systèmes *yeh-like* et *yde* qui sont majoritairement présents dans un état probablement non fonctionnel.

## 3.3 Identification des Loci d'Insertion

Nous avons utilisé l'alignement réalisé par progressiveMAUVE des 40 génomes pour comparer les sites d'insertion des différents systèmes. Nous avons trouvé que chaque groupe de systèmes orthologues est inséré au même locus dans les chromosomes, sauf pour les systèmes *sfa-foc* et *γ4-like* où deux sites d'insertion ont été observés. Ces deux systèmes sont localisés dans des PAI et ne sont présents que dans trois et deux génomes, respectivement. Le système *pap*, associé à un PAI et renfermant des paralogues, possède également trois sites d'insertion différents. L'unicité du site d'insertion pour les autres systèmes apporte une confirmation de leur lien de parenté (orthologues) et suggère un événement unique d'insertion au cours de l'évolution. Etonnamment, les systèmes *yqi* et *yqi-like* ont des sites d'insertion distants de seulement deux gènes.

## 3.4 Distribution des Répertoires des SystèmesCU

Parmi les génomes analysés, trois classes majeures de distribution des systèmes sont observées : i) les systèmes communs aux *E. coli*/*Shigella* et aux espèces proches qui devaient probablement être présents dans l'ancêtre commun à ces 60 génomes (*yad*, *yeh*, et *sfm*), ii) les systèmes présents seulement dans les souches de *E. coli*/*Shigella* et qui auraient été acquis, sous l'hypothèse d'une transmission verticale, avant la

divergence de ces souches (*yde*, *fim*, *ycb*, *yra*, *lpf1*, *lpf2 yeh-like* et *mat*), iii) les systèmes présentant une distribution sporadique sur l'arbre des espèces suggérant des acquisitions multiples et/ou récentes par transferts horizontaux (*auf*, *tsa*, *cupD*, *pap*, *γ2-like*, *γ4-like*, *yhc*, *sfa-foc*, *pix* et *K99*). Afin de mieux comprendre la trajectoire évolutive des systèmes CU et leur mode de diversification, nous avons modélisé les évènements de gain/perte de gènes sur l'arbre des espèces et recherché les évènements de recombinaison homologues et non-homologues qui auraient pu affecter les gènes de ces systèmes.

| Sous famille | Système | Organisation[#] | Complet | Altéré | Très altéré | Vestige | IS | Autre gène | AH | GC(ΔGC) |
|---|---|---|---|---|---|---|---|---|---|---|
| α | tsa | cpua | 12 | 1 | 4 | 2 | 1 | - | non (0/12) | 45,3 (-4,8) |
| | mat | pcuac | 26 | 1 | 2 | 1 | - | - | non (0/26) | 53,9 (+3,8) |
| β | yhc | pcu-IS-p | 5 | - | - | - | - | - | HGT (5/5) | 42 (-8,1) |
| γ1 | yde | pcuppa | 12 | 3 | 16 | 8 | - | - | HGT (11/12) | 44 (-6,1) |
| | yra | pcua | 17 | 4 | 8 | - | - | - | non (0/17) | 46,9 (-3,2) |
| | ycb | pcuappc | 16 | 7 | 8 | - | - | - | non (5/16) | 47,4 (-2,7) |
| | fim | ppcuppa | 25 | 3 | 7 | 1 | 1 | - | HGT (21/25) | 49 (-1,1) |
| | lpf1 | pcuap | 10 | 11 | 3 | - | - | - | non (1/6) | 46,3 (-3,8) |
| | lpf2 | pcua/pcuaa | 8 | 8 | 5 | 1 | 1 | - | non (3/8) | 45,4 (-4,7) |
| | sfa-foc | ppcuppa | 3 | - | - | - | - | - | PAI | 48,3 (-1,8) |
| | auf | pcuppca | 5 | 2 | 1 | - | 6 | 2 | HGT (4/4) | 40,6 (-9,5) |
| | sfm | pcuap | 38 | 3 | 8 | 2 | - | - | non (6/23) | 46 (-4,1) |
| γ2 | γ2-like | pccucpa | 1 | - | - | - | - | - | PAI | 42 (-8,1) |
| γ4 | yad | pcupppa | 40 | 8 | 6 | 4 | - | - | HGT (18/23) | 43,2 (-6,9) |
| | yeh | pcua | 44 | 7 | 4 | | 4 | 1 | HGT (23/25) | 41,7 (-8,4) |
| | yeh-like | pcua | 2 | 2 | 16 | 9 | 1 | 2 | HGT (2/2) | 37 (-13,1) |
| | cupD | ppcuac | 2 | 2 | - | - | - | - | PAI | 48 (-2,1) |
| | γ4-like | pcua | 2 | - | - | - | - | - | PAI | 42 (-8,1) |
| π | yfc | pucpppa | 26 | 9 | 5 | - | - | - | non (0/25) | 50,8 (+0,7) |
| | ybg | puca | 23 | 6 | 3 | 7 | | 1 | non (0/21) | 48,9 (-1,2) |
| | yqi | puca | 6 | - | - | - | - | - | non (0/5) | 44,8 (-5,3) |
| | yqi-like | puca | 12 | 6 | 3 | - | - | - | HGT (12/12) | 41,2 (-8,9) |
| | pap | ppucpppa | 8 | - | - | - | - | - | PAI | 48 (-2,1) |
| | pix | ppucppa | 1 | - | - | - | - | - | PAI | 48 (-2,1) |
| κ | K99 | puca | 1 | - | - | - | - | 25[&] | HGT (1/1) | 39 (-11,1) |

**Tableau 1**: Description des systèmes CU. Pour chaque groupe de systèmes sont indiqués : sa classification en sous-famille, son organisation chromosomique, le nombre de systèmes dans chacune des sept catégories de locus CU dans les 60 génomes, sa prédiction comme HGT par AlienHunter dans au moins 50% des génomes de *E. coli* codant pour des systèmes complets, sa localisation dans un îlot de pathogénicité (PAI), son taux de GC moyen dans les systèmes complets de *E. coli* et la différence par rapport au taux de GC moyen calculé sur les souches de *Escherichia/Shigella* (50,1%). [#]Chaque lettre indique le gène du système CU : « u » pour le usher, « c » pour la chaperonne, « p » pour la piline et « a » pour l'adhésine. [&]Présence des mêmes gènes *yibGJA-rhsA* au locus.

# 4 Dynamiques Evolutives des Systèmes CU

## 4.1 Identification des Transferts Horizontaux

La distribution sporadique sur l'arbre des espèces de certains groupes de systèmes suggère des acquisitions par transferts horizontaux. Des études précédentes ont montré que les systèmes *pap*, *sfa-foc*, *γ2-like*, *γ4-like* et *pix* sont localisés sur des îlots de pathogénicité (PAI I, II, III et V) dans les souches *E. coli 536* et *CFT073*. Notre analyse, basée sur Alien Hunter, a confirmé pour six systèmes (*cupD*, *pap*, *γ4-like*, *γ2-like*, *sfa-foc* et *pix*) leur présence dans des PAI et ceci pour tous les génomes où ils sont détectés.

Le taux moyen de GC des systèmes varie entre -13,1% et +3,8% par rapport au taux de GC moyen des souches *Escherichia/Shigella* (Table 1). Les 10 systèmes, *tsa*, *mat*, *yra*, *ycb*, *lpf1*, *lpf2*, *sfm*, *yfc*, *ybg* et *yqi* ne sont pas prédits comme HGT par Alien Hunter. Les neuf systèmes *yhc*, *yde*, *fim*, *auf*, *yad*, *yeh*, *yeh-like*, *yqi-like* et *K99* sont prédits acquis par HGT dans la majorité des génomes de *E. coli* ainsi que les six systèmes portés par des PAI. Ces résultats suggèrent que la majorité des systèmes CU sont acquis par HGT.

De plus, pour trois systèmes (*tsa*, *cupD,* et *γ4-like*) une similarité de séquence et une organisation génétique identique avec des systèmes portés par des plasmides [2] suggèrent une origine plasmidique.

## 4.2 Scénario d'Acquisition des Groupes de Systèmes CU

Nous avons retracé l'histoire évolutive des différents groupes de systèmes CU en utilisant une méthode de vraisemblance implémentée dans le logiciel Mesquite (cf. Matériel et Méthodes), en appliquant un modèle d'évolution à deux paramètres autorisant des taux différents pour le gain ou la perte de systèmes. Les six systèmes identifiés sur des PAI ont été exclus de l'analyse car acquis par HGT.

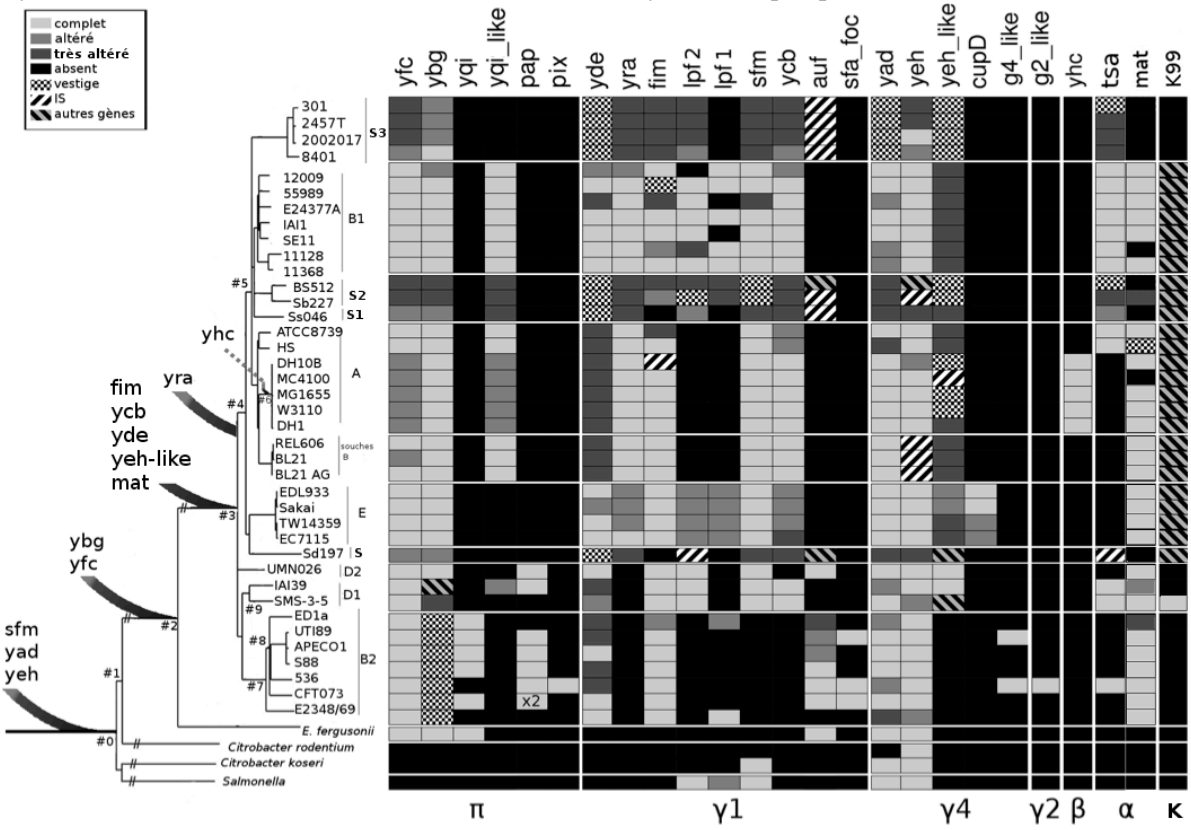



**Figure 1 : Description des répertoires de systèmes CU par rapport à l'évolution des souches *E.coli/Shigella***. Pour chaque souche de *E. coli,* le contenu de chaque locus CU est représenté en fonction de son appartenance à une des sept catégories comme décrit dans la légende en haut à gauche. Les différents phylogroupes sont mentionnés sur l'arbre des espèces. Pour les systèmes ayant été acquis une seule fois au cours de l'évolution (sauf *K99*), la branche correspondante est indiquée par une flèche. Les valeurs de bootstrap de chaque nœud sont supérieures à 70%, sauf pour la souche *E. coli UMN026* dont le placement reste ambiguë.

Parmi les 19 systèmes restants, 13 auraient été acquis une seule fois au cours de l'évolution (Figure 1 et Table 2) avec une forte probabilité de présence dans les nœuds profonds de l'arbre, excepté pour *yhc* et *K99*. Les systèmes *yad*, *yeh* et *sfm* auraient été acquis avant la divergence des espèces *Escherichia*/*Shigella*, *Salmonella* et *Citrobacter* (présence nœud #0). Les systèmes *ybg* et *yfc* auraient été acquis le long de la branche menant au dernier ancêtre commun (LCA) des *Escherichia*/*Shigella* (présence nœud #2), les quatre systèmes, *yde*, *fim*, *yeh-like* et *mat*, sur celle menant au LCA des souches de *E. coli*/*Shigella* (présence nœud #3) et le système *yra* sur celle menant au LCA des souches de *Shigella* et des souches appartenant aux phylogroupes A-souches B /B1 et E de *E. coli* (présence nœud #4). Pour le système *ycb*, deux scénarios sont envisageables, soit une acquisition sur la même branche que *yde*, *fim* et *yeh-like* (nœud #3) suivi de deux pertes, soit deux acquisitions indépendantes au même locus, l'une sur la branche menant au LCA des souches *IAI39* et *SMS-3-5* de *E. coli* (nœud #9) et l'autre sur la même branche que *yra* (nœud #4). Enfin, le système *yhc* présent uniquement dans les souches K-12, aurait été acquis le long de la branche menant à leur dernier ancêtre commun (présence nœud #6) et le système *K99* récemment dans la souche environnementale *SMS-3-5*. Les résultats de Mesquite montrent que les systèmes, après leur insertion dans un génome ancêtre, sont hérités verticalement. Pour les autres systèmes, le scénario le plus probable suggère plusieurs acquisitions indépendantes: deux acquisitions pour *lpf2*, *yqi* et *yqi-like* et trois acquisitions pour *auf* et *tsa*. La dynamique d'acquisition et de perte du système *lpf1* a été la plus difficile à appréhender. Il aurait été acquis cinq fois de façon indépendante au même locus chromosomique. Ces systèmes se distinguent des premiers par une plus grande labilité liée à des HGT multiples et/ou à des pertes complètes du système. De façon surprenante, ces

résultats montrent que des insertions indépendantes pourraient se produire au même locus. De plus, l'analyse du taux de GC des systèmes ne semble pas inversement proportionnel au temps passé du système dans le génome. D'une manière générale, l'annotation précise du contenu des locus CU permet d'observer qu'après leurs acquisitions, certains systèmes sont perdus (avec ou sans trace au locus) et parfois remplacés par des IS ou d'autres gènes (Table 2).

| Sous famille | Système | Congruence[&] (p-valeur test wsh) | Nombre d'acquisitions[#] | Nombre de pertes | Recombinaison homologue (feuille/nœud)[§] |
|---|---|---|---|---|---|
| α | tsa | NC (0) | 3 (#9: 0,806 ; #5: 0,925;1) | 1 | 3 (1/2) |
| | mat | NC (0) | 1 (#3 : 0,998) | 5 | 4 (3/1) |
| β | yhc | NA | 1 (#6: 0,999) | 0 | 0 |
| γ1 | yde | NC (0) | 1 (#3: 0,999) | 0 | 6 (4/2) |
| | yra | C (0,019) | 1 (#4: 0,999) | 0 | 0 |
| | ycb | C (0,302) | 1 (#3: 0,817) | 2 | 4 (0/4) |
| | fim | NC (0) | 1 (#3: 0,999) | 2 | 12 (10/2) |
| | lpf1 | NC (0) | 6 (0,749 à 1) | 5 | 2 (1/1) |
| | lpf2 | C[$] (0,270) | 2 (#3: 0,989 ; 0,999) | 3 | 1 (0/1) |
| | auf | NC (2.10-4) | 3 (#8: 0,999 ;1 ;1) | 0 | 2 (1/1) |
| | sfm | NC (0) | 1 (#0: 0,720) | 3 | 4 (3/1) |
| γ4 | yad | NC (0) | 1 (#0: 0,986) | 1 | 14 (8/6) |
| | yeh | NC (0) | 1 (#0: 0,990) | 2 | 2 (0/2) |
| | yeh-like | C (0,231) | 1 (#3: 0,946) | 4 | 0 |
| π | yfc | NC (0) | 1 (#2 : 0,936) | 0 | 15 (6/9) |
| | ybg | NC (0) | 1 (#2: 0,936) | 1 | 5 (2/3) |
| | yqi | C (0,021) | 2 (#8: 0,841 ;1) | 1 | 0 |
| | yqi-like | NC[%] (0,018) | 2 (#5: 0,943 ;1) | 1 | 2 (0/2) |

**Tableau 2 : Dynamiques évolutives des systèmes CU.** [&]Test de congruence avec CONSEL [20] entre l'arbre des systèmes et des espèces, NA (Non Applicable) au système *yhc* car trop peu de systèmes, NC (Non Congruent) indique que l'hypothèse H0 que les arbres sont congruents est rejetée avec pour seuil p-valeur $<10^{-3}$, C (Congruent) indique que l'hypothèse H0 de congruence des arbres ne peut être rejetée.[%] Mauvais positionnement de deux souches sur l'arbre du système confirmé par la recherche d'évènements de recombinaison.[#] Nombre de fois où le système aurait été acquis d'après Mesquite avec entre parenthèse le numéro du nœud se référant à l'arbre de la figure 1 et la probabilité de présence du système à ce nœud. [§] Nombre d'évènements de recombinaison prédit par ClonalOrigin avec entre parenthèse le nombre prédit aux feuilles et aux nœuds de la généalogie des souches. Les systèmes portés par des PAI. [$]Les arbres sont congruents mais la longueur des branches du phylogroupe E est plus grande qu'attendue.

## 4.3 Identification des Evènements de Recombinaison Homologue

Les arbres obtenus sur les séquences nucléotidiques des systèmes sont, en général, peu ou pas congruents avec l'arbre des espèces d'après le test wsh effectué par CONSEL [20] (Table 2) suggérant que les séquences ne sont pas issues d'un héritage strictement vertical. Des évènements de recombinaison homologue auraient pu avoir lieu après leur acquisition [21]. Nous avons testé cette hypothèse en utilisant ClonalOrigin (Table 2) qui permet de prédire les évènements de recombinaison homologue.

Pour les systèmes *yad*, *yfc* et *fim* dont les arbres des systèmes sont fortement incongruents avec celui des espèces, un très grand nombre de recombinaisons homologues (entre 12 et 15) a été prédit aux nœuds et aux feuilles de la généalogie des souches (Table 2) pouvant expliquer le mauvais positionnement d'un grand nombre de souches sur l'arbre du système. Il est de plus en accord avec des études précédentes identifiant les locus *fim* et *yfc* comme des points chauds de recombinaison [21]. En revanche, parmi les cinq systèmes ayant une topologie d'arbre congruente avec celle des espèces, aucun évènement de recombinaison n'a été détecté pour quatre d'entre eux (*yra*, *yhc*, *yqi* et *yeh-like*). Pour le cinquième, *ycb*, quatre évènements de recombinaison partielle ont été prédits qui n'ont pas affecté la topologie de l'arbre des espèces et n'ont pas non plus permis de trancher entre les deux scénarios envisagés ci-dessus. Dans les deux situations extrêmes analysées ci-dessus, les résultats obtenus sont ceux attendus (pas de recombinaison si congruence des arbres et inversement) et démontrent la capacité de ClonalOrigin à correctement détecter des évènements de recombinaison au locus CU. Pour les autres systèmes (sauf *sfm*), les incongruences d'arbres observées peuvent être expliquées par les évènements de recombinaison détectés aux nœuds et feuilles de la généalogie des souches. Trois classes d'évènements de recombinaison ont été observées : ceux affectant l'ensemble du système, ceux affectant environ 2/3 du système et ceux ne touchant que quelques gènes. Les évènements

partiels paraissent affecter les gènes du système de façon aléatoire, sans lien avec leur fonction.

## 5 Conclusion

La distribution des systèmes CU n'apparaît pas corrélée à la classification en pathovars des souches de *E. coli*. La majorité des systèmes semblent avoir été acquis une seule fois et avoir évolué par héritage vertical. Après insertion, ils ont pour la plupart subi différents évènements mutationnels (pseudogénéisation, recombinaison homologue) pouvant conduire à leur inactivation. Ce phénomène est particulièrement à l'œuvre dans les souches de *Shigella* où seuls deux systèmes sont conservés fonctionnels. La dynamique évolutive de ces systèmes chez les *Escherichia* pourrait être le reflet d'une adaptation rapide des souches à de nouveaux environnements. Les scénarios évolutifs les plus probables suggèrent que ces systèmes puissent s'intégrer et s'exciser de façon site spécifique, par un mécanisme moléculaire restant à élucider.

## Références


[1] B. C. Korea CG Ghigo JM, "The sweet connection: Solving the riddle of multiple sugar-binding fimbrial adhesins in Escherichia coli : Multiple E. coli fimbriae form a versatile arsenal of sugar-binding lectins potentially involved in surface-colonisation and tissue tropism.," *Bioessays*, vol. 33, pp. 300–311, 2011.

[2] D. J. Wurpel, S. a Beatson, M. Totsika, N. K. Petty, and M. a Schembri, "Chaperone-Usher Fimbriae of Escherichia coli.," *PloS one*, vol. 8, no. 1, p. e52835, Jan. 2013.

[3] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–402, Sep. 1997.

[4] K. Katoh, K. Kuma, H. Toh, and T. Miyata, "MAFFT version 5: improvement in accuracy of multiple sequence alignment.," *Nucleic acids research*, vol. 33, no. 2, pp. 511–8, Jan. 2005.

[5] S. Guindon and O. Gascuel, "A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, Oct. 2003.

[6] R. C. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity.," *BMC bioinformatics*, vol. 5, p. 113, Aug. 2004.

[7] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses," *Bioinformatics (Oxford, England)*, vol. 25, no. 15, Aug. 2009.

[8] F. Abascal, R. Zardoya, and D. Posada, "ProtTest: selection of best-fit models of protein evolution.," *Bioinformatics (Oxford, England)*, vol. 21, no. 9, pp. 2104–5, May 2005.

[9] P. J. Vernikos GS, "Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands.," *Bioinformatics*, vol. 22, pp. 2196–2203, 2006.

[10] M. G. I. Langille, W. W. L. Hsiao, and F. S. L. Brinkman, "Evaluation of genomic island predictors using a comparative genomics approach.," *BMC bioinformatics*, vol. 9, no. 1, p. 329, Jan. 2008.

[11] W. P. and D. R. M. Maddison, "Mesquite: a modular system for evolutionary analysis," 2011.

[12] X. Didelot, D. Lawson, A. Darling, and D. Falush, "Inference of homologous recombination in bacteria using whole-genome sequences.," *Genetics*, vol. 186, no. 4, pp. 1435–49, Dec. 2010.

[13] X. Didelot and D. Falush, "Inference of bacterial microevolution using multilocus sequence data.," *Genetics*, vol. 175, no. 3, pp. 1251–66, Mar. 2007.

[14] A. E. Darling, B. Mau, and N. T. Perna, "progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.," *PloS one*, vol. 5, no. 6, p. e11147, Jan. 2010.

[15] X. Didelot, G. Meric, D. Falush, A. E. Darling, and F. D. D. A. E. Didelot X Meric G, "Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli.," *BMC Genomics*, vol. 13, no. 1, p. 256, Jun. 2012.

[16] J. B. Kaper, J. P. Nataro, and H. L. Mobley, "Pathogenic Escherichia coli.," *Nature reviews. Microbiology*, vol. 2, no. 2, pp. 123–40, Feb. 2004.

[17] P. J. Herzer, S. Inouye, M. Inouye, and T. S. Whittam, "Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of Escherichia coli.," *Journal of bacteriology*, vol. 172, no. 11, pp. 6175–81, Nov. 1990.

[18] N. J. Holden, "Switches, cross-talk and memory in Escherichia coli adherence," *Journal of Medical Microbiology*, vol. 53, no. 7, pp. 585–593, Jul. 2004.

[19] B. A. J. Nuccio SP, "Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek.," *Microbiol Mol Biol Rev*, vol. 71, pp. 551–575, 2007.

[20] H. Shimodaira and M. Hasegawa, "CONSEL: for assessing the confidence of phylogenetic tree selection.," *Bioinformatics (Oxford, England)*, vol. 17, no. 12, pp. 1246–7, Dec. 2001.

[21] Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, et al., "Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths.," *PLoS Genet*, vol. 5, p. e1000344, 2009.

# Codon usage in *E. coli*: an evolutionary approach

Fanny POUYET[1], Julien JACQUEMETTON[1], Marc BAILLY-BECHET[1] and Laurent GUÉGUEN[1]

Lab. Biométrie et Biologie Évolutive, UMR5558 CNRS, 43 Bd du 11 nov. 1918, 69622 Villeurbane Cedex, France
{fanny.pouyet, julien.jacquemetton, marc.bailly-bechet, laurent.gueguen}@univ-lyon1.fr

**Abstract** *We develop a codon-based evolutionary model, based on previous works by Yang and Nielsen [11], with the capacity to distinguish between selective pressures acting specifically on codon usage or more generally on nucleotidic content. Our model, implemented in Bio++ [5] is multi-layered and allows to infer: i) the equilibrium frequencies for the nucleotidic mutational process; ii) the strengh of codon usage between the synonymous codons of each amino acid; and iii) the amino acid preferences. We apply this model in an homogeneous, non-stationary context on a dataset of three close* E. coli *strains [9] and show that codon usage and nucleotidic mutational process are counteracting each other, mutational process tending to increase dramatically the AT content in the equilibrium frequencies, while selection for codon usage acts towards a more balanced GC content.*

**Keywords** Evolutionary model, codon usage, mutational bias, *E. coli*.

## Usage du code chez *E. coli*: une approche évolutive

**Abstract** *Nous avons développé un modèle évolutif à l'échelle des codons, basé sur des travaux de Yang et Nielsen [11], qui permet de distinguer entre des pressions évolutives agissant sur l'usage du code ou sur la composition nucléotidique. Notre modèle, implémenté en Bio++ [5], est multi-couche et permet d'inférer: i) les fréquences d'équilibre du processus mutationnel à l'échelle nucléotidique, ii) l'intensité du biais d'usage du code entre codons synonymes pour chaque acide aminé, et iii) les préférences évolutives pour chaque acide aminé. Nous appliquons ce modèle à un jeu de données de gènes provenant de 3 souches proches de* E. coli *[9], et montrons que le biais d'usage du code génétique et les tendances mutationnelles à l'échelle nucléotidique sont en opposition: les processus mutationnels tendent vers une augmentation du contenu en AT des gènes, tandis que la sélection sur l'usage du code favorise un contenu plus équilibré en GC.*

**Keywords** Modèle évolutif, usage du code, biais mutationnels, *E. coli*

The genetic code is degenerated, allowing for different codons – said nsynonymous – to code for the same amino acid. Codon usage bias is the non-random preference, in a gene or more generally a genome, for using a particular codon over synonymous ones. Thanks to the accumulation of genomic sequences, codon bias has been documented in all living organisms – see [6,10] for a review. Various causes have been advanced to explain the existence of this bias; amongst them, two main hypotheses have emerged: *mutational bias* and *translational selection*. The *mutational bias* hypothesis postulates that codon usage is due to global or local biases in the mutation patterns, that affect in particular unselected positions in a sequence, such as the third position of codons in amino acids four times degenerated, where all changes have no consequence on the protein sequence. Conversely, the *translational selection* hypothesis says that the choice of a particular codon can benefit the organism, by making the translation of this codon – and more generally the translation of the entire gene – more accurate or more efficient. This hypothesis is supported by the observed correlation between codon frequencies and cognate tRNA frequencies in various bacteria [3], correlation that has been predicted in models of translational selection [2].

In bacterial genomes, both forces seem to act, with translational selection being very strong mainly for highly expressed proteins such as ribosomal proteins. In order to disentangle these two forces, we developed an evolutionary model, inspired from the works of Yang and Nielsen [11], in which we can infer a preference parameter for each codon, relative to its synonymous ones. These preferences, denoted $\phi_{aa}(i)$ for codon $i$ inside

amino acid $aa$, are the relative equilibrium frequencies of the codons, inside their amino acids, if only selection for codon usage was acting on the sequences. Moreover, our model includes classical equilibrium nucleotidic frequencies to account for mutational biases. In a more mathematical way, one can write the generator $q_{ij}$ of our model, i.e. the probability to observe a substitution between codon $i$ and codon $j$, as:

$$q_{ij} \propto \begin{cases} 0 & \text{if more than one nucleotide change} \\ \pi_{j_p}.\frac{-\log\left(\frac{\phi_{aa}(i)}{\phi_{aa}(j)}\right)}{1-\frac{\phi_{aa}(i)}{\phi_{aa}(j)}} & \text{synonymous transversion} \\ \pi_{j_p}.\kappa.\frac{-\log\left(\frac{\phi_{aa}(i)}{\phi_{aa}(j)}\right)}{1-\frac{\phi_{aa}(i)}{\phi_{aa}(j)}} & \text{synonymous transition} \\ \pi_{j_p}.\omega.\frac{-\log\left(\frac{\psi_{aa_i}\phi_{aa_i}(i)}{\psi_{aa_j}\phi_{aa_j}(j)}\right)}{1-\frac{\psi_{aa_i}\phi_{aa_i}(i)}{\psi_{aa_j}\phi_{aa_j}(j)}} & \text{non synonymous transversion} \\ \pi_{j_p}.\kappa.\omega.\frac{-\log\left(\frac{\psi_{aa_i}\phi_{aa_i}(i)}{\psi_{aa_j}\phi_{aa_j}(j)}\right)}{1-\frac{\psi_{aa_i}\phi_{aa_i}(i)}{\psi_{aa_j}\phi_{aa_j}(j)}} & \text{non synonymous transition} \end{cases},$$

with $\pi_{j_p}$ being the equilibrium frequency of nucleotide $j_p$ ($p$ indicating the position in codon $j$), $\kappa$ the transition/transversion ratio, $\omega$ the non-synonymous/synonymous ratio, and $\psi_{aa_j}$ a preference towards amino acid $j$ relative to other amino acids. We then have a three-layered model, with:

– the nucleotidic layer with parameters $\pi_{j_p}$ and $\kappa$,
– the codon usage layer with parameters $\phi_{aa}(i)$,
– the amino acid layer with parameters $\psi_{aa_j}$ and $\omega$.

This formulation is a reparametrization of the previous works of [11]; the main difference in our version of the model is the complete separation between the codon usage parameters and amino acid parameters, ensuring that pressures acting on both biological scales can be measured separately. Then, we have 67 parameters: 5 for the nucleotidic layer, 41 for the codon layer and 21 for the amino acid layer. This is a quite high number, that can however be dealt with, as shown in [11]. All these three layers can be infered jointly from the data, and given a set of parameters, one can compute the equilibrium frequencies of all codons (and then nucleotides or amino acids) under the effect of all three of them, or only one by one. Computationally, this model has been implemented in the Bio++ suite [5].

The model was applied to a dataset composed of 3353 orthologuous genes in 3 closely related strains of *E. coli*, coming from [9], in order to study the relative importance of mutational biases and codon preferences. As a single gene alignment does not contain enough information to estimate all parameters of the model, we built gene concatenates in the following way. Genes were first sorted by increasing Fop values; Fop is the fraction of optimal codons in a gene, one of the common measure of codon bias strength [8]. The gene list was segmented in 33 consecutive groups of 102 genes each (except for the last one, 89 genes), and all genes in the group were concatenated to be studied as a single sequence. By definition of the Fop, the last concatenate mostly contains ribosomal proteins and factors involved in translation, *i.e* highly expressed genes.

Another way to group genes, instead of using Fop values measuring codon usage intensity, would have been to cluster together genes sharing the same codon bias, using e.g. the method by [1]; here we preferred grouping by Fop in order to keep close to the standard literature of codon usage bias in bacteria and make comparisons easy. Descriptive statistics of our 33 concatenates can be found in TABLE 1.

In this paper, we will only focus on the analysis of the codon and nucleotidic parameters, the amino acid content evolution in these three closely related strains being negligible in first approach. The obtained parameters are reported in TABLE 1. We first verified that inferring codon usage parameters did not modify strongly the value of known parameters, such as $\omega$. Indeed, it is known that, due to purifying selection, $\omega$ values tend to decrease in genes with strong codon usage bias (and a high Fop). We checked that this was still true in our analyses: indeed, the concatenate mean Fop and $\omega$ correlates fairly well (Spearman $\rho = -0.782$, p$< 10^{-7}$).

Then, we looked for the values obtained for codon preferences. As one can see in FIG. 1, these values are quite consistent from one concatenate to another, often giving the same trend of the entire genome: as

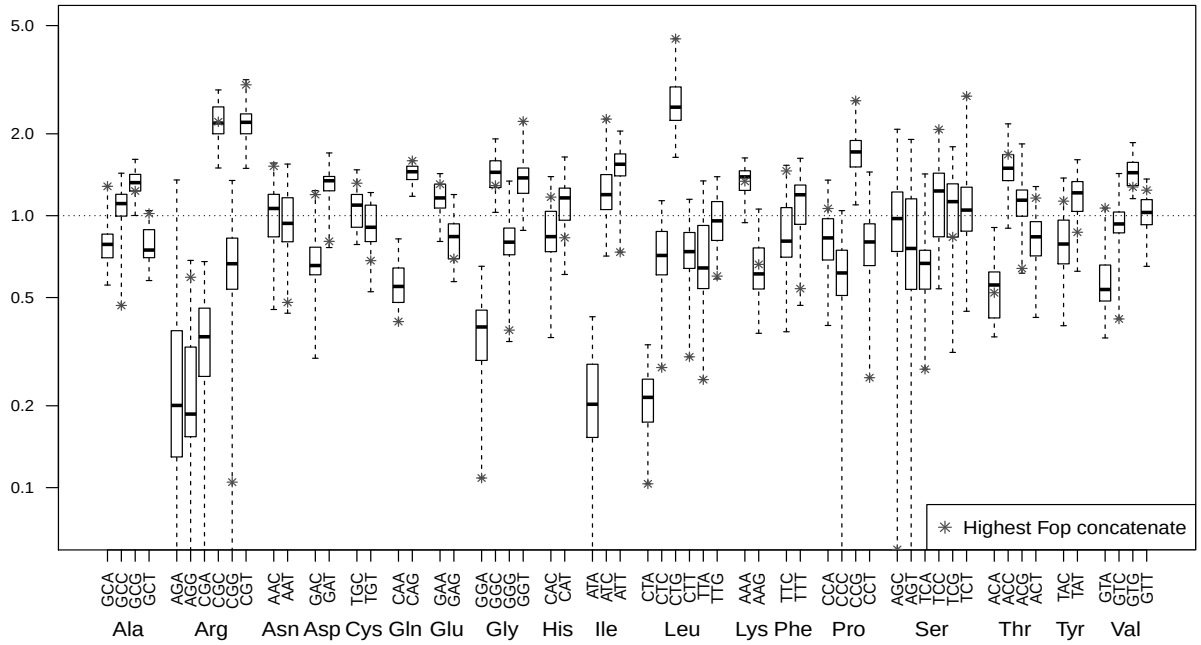



**Figure 1.** Codon preferences; $x$-axis, the codons, organized by amino acid; $y$-axis, the normalized preferences, ie preferences multiplied by the degeneracy of the amino acid, such that a normalized preference of 1 means uniform preferences for synonymous codons in the amino acid. The boxplot extends to the extreme values; codon preferences in the highest Fop concatenate, containing ribosomal proteins, are shown by a star.

an example, for Glutamine, codon CAG is always prefered over codon CAA. Note that this is not a trival consequence of the nucleotidic content of the amino acid, as in Lysine, coded for by AAA and AAG, the A-ending codon is almost always prefered over the G-ending one. In some cases, such as Serine, preferences are much more volatile, with no clear-cut rule. This confirms *a posteriori* one of the potentiel interests of the model, *i. e.* to be able to discriminate on which amino acids selection on codon usage is visible, and to go from a single measure of codon bias intensity, such as Fop or CAI, to a more detailed, amino-acid dependent characterization. Finally, one can see on this graph that the concatenate including genes with the highest Fop, ie. mostly ribosomal proteins, often shows a more marked preference towards a codon relative to the others, as indicated by the extreme star positions in the boxplots: see by example the Leucine or Isoleucine plots. This marked preference is in agreement with what was expected for these genes, which are used as a reference for the codon usage bias of the whole organism in measures like Fop.

Finally, we studied the GC content at equilibrium (GC*) in the 33 concatenates that would result from selection on codon usage only, or mutational bias only. In reality, both layers influence sequence evolution, and global equilibrium frequencies in this dataset are much closer to those given by the nucleotidic layer than by the codon usage layer (data not shown); but, aside from the question of the relative strength of both pressures on equilibrium frequencies – which is still under study – one can ask how these two pressures relate. As shown in FIG. 2, the codon and nucleotidic layers are counteracting each other, the mutational bias decreasing dramatically the GC content, while the codon layer tends towards a more balanced GC*. Both these facts could be independently supposed, as a global mutational bias towards AT has been documented in bacteria [7], and the codon layer, by definition, can not have a strong effect on the average nucleotide composition; but the negative correlation observed, implying that stronger mutational biases are countered by stronger selection on codon usage, was unexpected. Moreover, one can note on this graph that genes with a stronger Fop (darker points) tend to cluster above the regression line, showing that they are significantly more affected by the codon layer than the low Fop genes (Pearson correlation test between Fop and difference between GC* of the codon layer and the regression line , $R = 0.79$, $p < 2.10^{-8}$). These results implicate that selection on codon usage is higher on sequences submitted to high mutational biases; one could hypothesize that codon usage bias is then

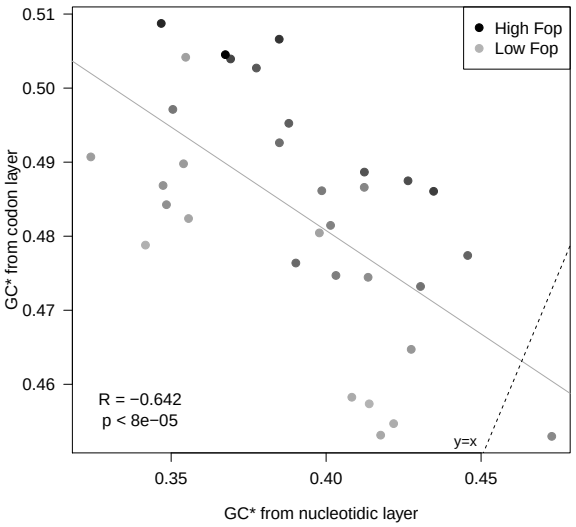



**Figure 2.** Comparison of GC* from the nucleotidic ($x$-axis) and codon ($y$-axis) layers. The Spearman correlation coefficient and p-value of the corresponding test are given. Note the difference in range between the two axis. The $y = x$ line is shown in dots to locate where the points should align if the GC* of both layers were the same.

a way to prevent a fast sequence degradation due to AT biased mutational processes. A next step in this study could be to analyze the set of genes for which both evolutive layers tend to close values (points on bottom right of FIG. 2), and those which are far from it (top left), and see if any qualitative differences exist between them.

Thus, we have developed a codon-based, multi-layered evolutionary model, for the study of codon usage bias, in the Bio++ suite. Our first analyses, on a simple dataset, show that our results are reliable and allow an easy intepretation of the different forces acting on sequence evolution, by example in shedding light on the antagonists effects of mutational bias and selection for codon usage in *E. coli*. These tools could be adapted and used in various phylogenetic contexts, to help untangle evolutionary effects that can hardly be distinguished based on sequence analysis alone.

## References

[1] M. Bailly-Bechet, A. Danchin, M. Iqbal, M. Marsili and M. Vergassola. Codon usage domains over bacterial chromosomes. *PLoS Comp. Biol.*, 2(4):e37, 2006.

[2] M. Bulmer. Coevolution of codon usage and transfer rna abundance. *Nature*, 325(6106):728–730, 1987.

[3] H. Dong, L. Nilsson, and C. G. Kurland. Co-variation of trna abundance and codon usage in escherichia coli at different growth rates. *J Mol Biol*, 260(5):649–663, Aug 1996.

[4] J. Dutheil, S. Gaillard, E. Bazin, S. Glémin, V. Ranwez, N. Galtier, and K. Belkhir. Bio++: a set of c++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, 7:188, 2006.

[5] L. Guéguen, S. Gaillard, B. Boussau, M. Gouy, M. Groussin, N.C. Rochette, T. Bigot, D. Fournier, F. Pouyet, V. Cahais, A. Bernard, C. Scornavacca, B. Nabholz, A. Haudry, L. Dachary, N. Galtier, K. Belkhir, J.Y. Dutheil Bio++: efficient, extensible libraries and tools for computational molecular evolution *Mol Biol Evol*, 2013.

[6] R. Hershberg and D. A. Petrov. Selection on codon bias. *Annu Rev Genet*, 42:287–299, 2008.

[7] R. Hershberg and D. A. Petrov. Evidence that mutation is universally biased towards at in bacteria. *PLoS Genet*, 6(9), Sep 2010.

[8] T. Ikemura. Correlation between the abundance of *Escherichia coli* tRNAs and the occurence of the respective codons in its protein genes. *J Mol Biol* 146:1–21

[9] I. K. Jordan, F. A. Kondrashov, I. A. Adzhubei, Y. I. Wolf, E. V. Koonin, A. S. Kondrashov, and S. Sunyaev. A universal trend of amino acid gain and loss in protein evolution. *Nature*, 433(7026):633–638, Feb 2005.

[10] P. M. Sharp, L. R. Emery, and K. Zeng. Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci*, 365:1203–1212, Apr 2010.

[11] Z. Yang and R. Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*, 25(3):568–579, Mar 2008.

| **Concatenate** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (kb) | 212.68 | 232.31 | 255.00 | 276.19 | 285.91 | 289.30 | 286.07 | 300.20 | 279.43 | 280.17 | 306.01 |
| GC% | 0.48 | 0.51 | 0.50 | 0.51 | 0.51 | 0.51 | 0.52 | 0.52 | 0.52 | 0.52 | 0.52 |
| Identity% | 0.83 | 0.90 | 0.84 | 0.91 | 0.88 | 0.87 | 0.92 | 0.92 | 0.90 | 0.90 | 0.92 |
| Mean Fop | 0.34 | 0.35 | 0.36 | 0.37 | 0.38 | 0.38 | 0.39 | 0.39 | 0.40 | 0.40 | 0.41 |
| **Estimated Parameters** | | | | | | | | | | | |
| $\omega$ | 0.27 | 0.11 | 0.18 | 0.14 | 0.18 | 0.10 | 0.14 | 0.14 | 0.08 | 0.12 | 0.13 |
| $\pi_A$ | 0.34 | 0.38 | 0.36 | 0.35 | 0.36 | 0.35 | 0.33 | 0.36 | 0.43 | 0.41 | 0.32 |
| $\pi_C$ | 0.22 | 0.16 | 0.24 | 0.21 | 0.21 | 0.17 | 0.18 | 0.17 | 0.17 | 0.20 | 0.22 |
| $\pi_G$ | 0.20 | 0.18 | 0.18 | 0.20 | 0.21 | 0.18 | 0.21 | 0.18 | 0.15 | 0.15 | 0.21 |
| $\pi_T$ | 0.25 | 0.28 | 0.22 | 0.25 | 0.22 | 0.30 | 0.27 | 0.29 | 0.24 | 0.24 | 0.25 |
| Codon preferences $\phi_{aa}(i)$ | | | | | | | | | | | |
| GCA | 0.21 | 0.18 | 0.17 | 0.19 | 0.18 | 0.20 | 0.21 | 0.17 | 0.17 | 0.16 | 0.20 |
| GCC | 0.27 | 0.32 | 0.22 | 0.23 | 0.28 | 0.30 | 0.29 | 0.32 | 0.31 | 0.25 | 0.26 |
| GCG | 0.38 | 0.30 | 0.40 | 0.31 | 0.31 | 0.32 | 0.33 | 0.36 | 0.35 | 0.34 | 0.32 |
| GCT | 0.14 | 0.21 | 0.21 | 0.26 | 0.23 | 0.18 | 0.16 | 0.15 | 0.16 | 0.24 | 0.22 |
| AGA | 0.07 | 0.03 | 0.03 | 0.06 | 0.04 | 0.06 | 0.06 | 0.03 | 0.03 | 0.01 | 0.06 |
| AGG | 0.08 | 0.05 | 0.04 | 0.08 | 0.07 | 0.04 | 0.03 | 0.02 | 0.03 | 0.03 | 0.05 |
| CGA | 0.08 | 0.07 | 0.06 | 0.09 | 0.06 | 0.06 | 0.10 | 0.08 | 0.08 | 0.04 | 0.07 |
| CGC | 0.34 | 0.43 | 0.41 | 0.30 | 0.33 | 0.40 | 0.36 | 0.48 | 0.33 | 0.33 | 0.33 |
| CGG | 0.18 | 0.16 | 0.10 | 0.13 | 0.12 | 0.15 | 0.12 | 0.13 | 0.19 | 0.22 | 0.14 |
| CGT | 0.25 | 0.26 | 0.36 | 0.35 | 0.38 | 0.29 | 0.33 | 0.26 | 0.34 | 0.36 | 0.34 |
| AAC | 0.24 | 0.42 | 0.23 | 0.31 | 0.32 | 0.52 | 0.57 | 0.62 | 0.43 | 0.47 | 0.39 |
| AAT | 0.76 | 0.58 | 0.77 | 0.69 | 0.68 | 0.48 | 0.43 | 0.38 | 0.57 | 0.53 | 0.61 |
| GAC | 0.20 | 0.32 | 0.15 | 0.24 | 0.20 | 0.37 | 0.30 | 0.49 | 0.31 | 0.30 | 0.31 |
| GAT | 0.80 | 0.68 | 0.85 | 0.76 | 0.80 | 0.63 | 0.70 | 0.51 | 0.69 | 0.70 | 0.69 |
| TGC | 0.40 | 0.39 | 0.44 | 0.43 | 0.50 | 0.45 | 0.62 | 0.57 | 0.60 | 0.65 | 0.44 |
| TGT | 0.60 | 0.61 | 0.56 | 0.57 | 0.50 | 0.55 | 0.38 | 0.43 | 0.40 | 0.35 | 0.56 |
| CAA | 0.37 | 0.38 | 0.32 | 0.37 | 0.36 | 0.26 | 0.41 | 0.30 | 0.26 | 0.23 | 0.35 |
| CAG | 0.63 | 0.62 | 0.68 | 0.63 | 0.64 | 0.74 | 0.59 | 0.70 | 0.74 | 0.77 | 0.65 |
| GAA | 0.48 | 0.40 | 0.52 | 0.53 | 0.61 | 0.54 | 0.54 | 0.52 | 0.47 | 0.42 | 0.63 |
| GAG | 0.52 | 0.60 | 0.48 | 0.47 | 0.39 | 0.46 | 0.46 | 0.48 | 0.53 | 0.58 | 0.37 |
| GGA | 0.15 | 0.15 | 0.11 | 0.15 | 0.11 | 0.11 | 0.11 | 0.09 | 0.07 | 0.07 | 0.11 |
| GGC | 0.36 | 0.28 | 0.42 | 0.26 | 0.32 | 0.38 | 0.43 | 0.40 | 0.33 | 0.27 | 0.31 |
| GGG | 0.27 | 0.33 | 0.18 | 0.22 | 0.20 | 0.18 | 0.19 | 0.21 | 0.30 | 0.34 | 0.18 |
| GGT | 0.22 | 0.24 | 0.29 | 0.37 | 0.37 | 0.32 | 0.27 | 0.30 | 0.30 | 0.33 | 0.40 |
| CAC | 0.27 | 0.25 | 0.18 | 0.37 | 0.25 | 0.52 | 0.30 | 0.46 | 0.44 | 0.37 | 0.37 |
| CAT | 0.73 | 0.75 | 0.82 | 0.63 | 0.75 | 0.48 | 0.70 | 0.54 | 0.56 | 0.63 | 0.63 |
| ATA | 0.12 | 0.09 | 0.06 | 0.14 | 0.10 | 0.12 | 0.13 | 0.08 | 0.07 | 0.06 | 0.06 |
| ATC | 0.24 | 0.40 | 0.26 | 0.32 | 0.30 | 0.37 | 0.32 | 0.45 | 0.40 | 0.33 | 0.36 |
| ATT | 0.65 | 0.51 | 0.68 | 0.54 | 0.60 | 0.52 | 0.54 | 0.47 | 0.53 | 0.60 | 0.58 |
| CTA | 0.03 | 0.05 | 0.02 | 0.04 | 0.03 | 0.04 | 0.04 | 0.03 | 0.02 | 0.03 | 0.05 |
| CTC | 0.10 | 0.16 | 0.08 | 0.13 | 0.09 | 0.18 | 0.12 | 0.17 | 0.19 | 0.10 | 0.08 |
| CTG | 0.28 | 0.36 | 0.28 | 0.27 | 0.30 | 0.34 | 0.42 | 0.44 | 0.38 | 0.42 | 0.39 |
| CTT | 0.14 | 0.17 | 0.18 | 0.16 | 0.19 | 0.19 | 0.13 | 0.12 | 0.12 | 0.13 | 0.12 |
| TTA | 0.22 | 0.12 | 0.22 | 0.22 | 0.18 | 0.10 | 0.11 | 0.09 | 0.09 | 0.09 | 0.19 |
| TTG | 0.23 | 0.15 | 0.22 | 0.18 | 0.21 | 0.15 | 0.19 | 0.16 | 0.19 | 0.23 | 0.17 |
| AAA | 0.58 | 0.58 | 0.60 | 0.57 | 0.71 | 0.68 | 0.64 | 0.65 | 0.52 | 0.47 | 0.70 |
| AAG | 0.42 | 0.42 | 0.40 | 0.43 | 0.29 | 0.32 | 0.36 | 0.35 | 0.48 | 0.53 | 0.30 |
| TTC | 0.26 | 0.36 | 0.19 | 0.24 | 0.23 | 0.36 | 0.35 | 0.47 | 0.40 | 0.32 | 0.30 |
| TTT | 0.74 | 0.64 | 0.81 | 0.76 | 0.77 | 0.64 | 0.65 | 0.53 | 0.60 | 0.68 | 0.70 |
| CCA | 0.17 | 0.16 | 0.14 | 0.21 | 0.17 | 0.20 | 0.24 | 0.14 | 0.10 | 0.16 | 0.20 |
| CCC | 0.17 | 0.26 | 0.16 | 0.19 | 0.22 | 0.24 | 0.17 | 0.26 | 0.22 | 0.18 | 0.17 |
| CCG | 0.37 | 0.35 | 0.38 | 0.32 | 0.27 | 0.39 | 0.42 | 0.46 | 0.45 | 0.43 | 0.35 |
| CCT | 0.29 | 0.23 | 0.33 | 0.29 | 0.34 | 0.17 | 0.16 | 0.14 | 0.23 | 0.23 | 0.28 |
| AGC | 0.16 | 0.27 | 0.19 | 0.12 | 0.20 | 0.17 | 0.22 | 0.14 | 0.12 | 0.16 | 0.14 |
| AGT | 0.22 | 0.32 | 0.27 | 0.17 | 0.30 | 0.14 | 0.12 | 0.13 | 0.11 | 0.19 | 0.09 |
| TCA | 0.19 | 0.10 | 0.11 | 0.12 | 0.09 | 0.15 | 0.12 | 0.12 | 0.12 | 0.09 | 0.15 |
| TCC | 0.12 | 0.09 | 0.10 | 0.23 | 0.10 | 0.22 | 0.19 | 0.24 | 0.27 | 0.12 | 0.16 |
| TCG | 0.16 | 0.15 | 0.22 | 0.19 | 0.13 | 0.17 | 0.21 | 0.21 | 0.22 | 0.27 | 0.28 |
| TCT | 0.15 | 0.07 | 0.12 | 0.17 | 0.18 | 0.16 | 0.12 | 0.16 | 0.17 | 0.18 | 0.18 |
| ACA | 0.21 | 0.17 | 0.14 | 0.15 | 0.16 | 0.21 | 0.11 | 0.17 | 0.10 | 0.10 | 0.16 |
| ACC | 0.26 | 0.35 | 0.22 | 0.34 | 0.32 | 0.27 | 0.43 | 0.36 | 0.33 | 0.24 | 0.32 |
| ACG | 0.34 | 0.31 | 0.36 | 0.28 | 0.32 | 0.31 | 0.29 | 0.36 | 0.31 | 0.46 | 0.33 |
| ACT | 0.18 | 0.17 | 0.27 | 0.24 | 0.21 | 0.21 | 0.17 | 0.11 | 0.26 | 0.20 | 0.19 |
| TAC | 0.20 | 0.36 | 0.20 | 0.26 | 0.22 | 0.31 | 0.32 | 0.53 | 0.32 | 0.37 | 0.34 |
| TAT | 0.80 | 0.64 | 0.80 | 0.74 | 0.78 | 0.69 | 0.68 | 0.47 | 0.68 | 0.63 | 0.66 |
| GTA | 0.13 | 0.10 | 0.10 | 0.13 | 0.14 | 0.13 | 0.16 | 0.11 | 0.09 | 0.09 | 0.17 |
| GTC | 0.22 | 0.36 | 0.21 | 0.26 | 0.23 | 0.25 | 0.25 | 0.26 | 0.27 | 0.23 | 0.22 |
| GTG | 0.39 | 0.29 | 0.37 | 0.29 | 0.33 | 0.41 | 0.36 | 0.46 | 0.36 | 0.44 | 0.32 |
| GTT | 0.26 | 0.25 | 0.32 | 0.33 | 0.30 | 0.20 | 0.23 | 0.16 | 0.28 | 0.24 | 0.29 |

| **Concatenate** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **21** | **22** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (kb) | 295.73 | 315.13 | 298.40 | 291.13 | 272.28 | 290.82 | 337.11 | 267.95 | 291.40 | 315.53 | 344.65 |
| GC% | 0.52 | 0.53 | 0.52 | 0.53 | 0.52 | 0.53 | 0.53 | 0.53 | 0.53 | 0.53 | 0.54 |
| Identity% | 0.94 | 0.93 | 0.93 | 0.91 | 0.92 | 0.94 | 0.94 | 0.94 | 0.91 | 0.93 | 0.94 |
| Mean Fop | 0.41 | 0.42 | 0.42 | 0.43 | 0.43 | 0.43 | 0.44 | 0.44 | 0.45 | 0.45 | 0.46 |
| **Estimated Parameters** | | | | | | | | | | | |
| $\omega$ | 0.10 | 0.08 | 0.13 | 0.11 | 0.11 | 0.09 | 0.09 | 0.05 | 0.04 | 0.07 | 0.06 |
| $\pi_A$ | 0.40 | 0.40 | 0.32 | 0.28 | 0.33 | 0.33 | 0.34 | 0.30 | 0.33 | 0.31 | 0.34 |
| $\pi_C$ | 0.18 | 0.19 | 0.20 | 0.24 | 0.19 | 0.20 | 0.21 | 0.19 | 0.17 | 0.23 | 0.19 |
| $\pi_G$ | 0.17 | 0.16 | 0.21 | 0.23 | 0.22 | 0.20 | 0.19 | 0.21 | 0.18 | 0.20 | 0.19 |
| $\pi_T$ | 0.26 | 0.25 | 0.26 | 0.25 | 0.26 | 0.27 | 0.25 | 0.30 | 0.31 | 0.26 | 0.27 |
| Codon preferences $\phi_{aa}(i)$ | | | | | | | | | | | |
| GCA | 0.16 | 0.14 | 0.18 | 0.25 | 0.20 | 0.19 | 0.18 | 0.23 | 0.14 | 0.21 | 0.18 |
| GCC | 0.28 | 0.29 | 0.31 | 0.26 | 0.29 | 0.24 | 0.25 | 0.21 | 0.36 | 0.25 | 0.32 |
| GCG | 0.37 | 0.39 | 0.29 | 0.31 | 0.34 | 0.34 | 0.33 | 0.38 | 0.33 | 0.36 | 0.34 |
| GCT | 0.20 | 0.18 | 0.21 | 0.19 | 0.17 | 0.22 | 0.25 | 0.18 | 0.18 | 0.18 | 0.16 |
| AGA | 0.03 | 0.01 | 0.05 | 0.11 | 0.01 | 0.04 | 0.05 | 0.07 | 0.01 | 0.07 | 0.03 |
| AGG | 0.04 | 0.01 | 0.02 | 0.01 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.02 | 0.01 |
| CGA | 0.04 | 0.06 | 0.09 | 0.11 | 0.05 | 0.05 | 0.04 | 0.08 | 0.06 | 0.07 | 0.04 |
| CGC | 0.45 | 0.39 | 0.35 | 0.25 | 0.43 | 0.35 | 0.32 | 0.37 | 0.42 | 0.37 | 0.46 |
| CGG | 0.14 | 0.15 | 0.11 | 0.10 | 0.11 | 0.12 | 0.14 | 0.14 | 0.11 | 0.10 | 0.09 |
| CGT | 0.30 | 0.38 | 0.39 | 0.42 | 0.38 | 0.41 | 0.42 | 0.30 | 0.37 | 0.37 | 0.38 |
| AAC | 0.55 | 0.42 | 0.46 | 0.35 | 0.59 | 0.39 | 0.45 | 0.69 | 0.56 | 0.41 | 0.57 |
| AAT | 0.45 | 0.58 | 0.54 | 0.65 | 0.41 | 0.61 | 0.55 | 0.31 | 0.44 | 0.59 | 0.43 |
| GAC | 0.33 | 0.27 | 0.35 | 0.23 | 0.33 | 0.35 | 0.24 | 0.31 | 0.32 | 0.29 | 0.38 |
| GAT | 0.67 | 0.73 | 0.65 | 0.77 | 0.67 | 0.65 | 0.76 | 0.69 | 0.68 | 0.71 | 0.62 |
| TGC | 0.55 | 0.66 | 0.52 | 0.44 | 0.59 | 0.52 | 0.49 | 0.42 | 0.55 | 0.59 | 0.64 |
| TGT | 0.45 | 0.34 | 0.48 | 0.56 | 0.41 | 0.48 | 0.51 | 0.58 | 0.45 | 0.41 | 0.36 |
| CAA | 0.24 | 0.21 | 0.29 | 0.41 | 0.27 | 0.27 | 0.28 | 0.29 | 0.21 | 0.32 | 0.25 |
| CAG | 0.76 | 0.79 | 0.71 | 0.59 | 0.73 | 0.73 | 0.72 | 0.71 | 0.79 | 0.68 | 0.75 |
| GAA | 0.47 | 0.53 | 0.58 | 0.66 | 0.67 | 0.60 | 0.56 | 0.59 | 0.61 | 0.59 | 0.58 |
| GAG | 0.53 | 0.47 | 0.42 | 0.34 | 0.33 | 0.40 | 0.44 | 0.41 | 0.39 | 0.41 | 0.42 |
| GGA | 0.07 | 0.11 | 0.16 | 0.14 | 0.12 | 0.13 | 0.09 | 0.08 | 0.07 | 0.10 | 0.08 |
| GGC | 0.37 | 0.31 | 0.31 | 0.28 | 0.40 | 0.36 | 0.34 | 0.32 | 0.44 | 0.30 | 0.41 |
| GGG | 0.22 | 0.26 | 0.23 | 0.23 | 0.20 | 0.20 | 0.20 | 0.21 | 0.18 | 0.20 | 0.14 |
| GGT | 0.34 | 0.32 | 0.30 | 0.34 | 0.28 | 0.32 | 0.38 | 0.38 | 0.31 | 0.39 | 0.37 |
| CAC | 0.39 | 0.30 | 0.40 | 0.36 | 0.42 | 0.51 | 0.37 | 0.46 | 0.42 | 0.40 | 0.49 |
| CAT | 0.61 | 0.70 | 0.60 | 0.64 | 0.58 | 0.49 | 0.63 | 0.54 | 0.58 | 0.60 | 0.51 |
| ATA | 0.11 | 0.05 | 0.09 | 0.09 | 0.08 | 0.06 | 0.10 | 0.08 | 0.05 | 0.14 | 0.04 |
| ATC | 0.39 | 0.35 | 0.38 | 0.35 | 0.44 | 0.41 | 0.35 | 0.44 | 0.55 | 0.38 | 0.43 |
| ATT | 0.50 | 0.60 | 0.52 | 0.56 | 0.48 | 0.53 | 0.55 | 0.48 | 0.39 | 0.48 | 0.53 |
| CTA | 0.03 | 0.03 | 0.04 | 0.05 | 0.04 | 0.04 | 0.04 | 0.04 | 0.06 | 0.04 | 0.04 |
| CTC | 0.11 | 0.11 | 0.12 | 0.09 | 0.15 | 0.09 | 0.12 | 0.11 | 0.12 | 0.10 | 0.16 |
| CTG | 0.37 | 0.42 | 0.37 | 0.31 | 0.38 | 0.50 | 0.40 | 0.51 | 0.49 | 0.39 | 0.47 |
| CTT | 0.12 | 0.16 | 0.13 | 0.12 | 0.17 | 0.10 | 0.14 | 0.07 | 0.11 | 0.12 | 0.12 |
| TTA | 0.15 | 0.10 | 0.16 | 0.21 | 0.11 | 0.10 | 0.11 | 0.12 | 0.09 | 0.20 | 0.08 |
| TTG | 0.20 | 0.18 | 0.17 | 0.21 | 0.15 | 0.17 | 0.19 | 0.14 | 0.13 | 0.14 | 0.14 |
| AAA | 0.62 | 0.59 | 0.75 | 0.73 | 0.70 | 0.73 | 0.70 | 0.79 | 0.66 | 0.76 | 0.75 |
| AAG | 0.38 | 0.41 | 0.25 | 0.27 | 0.30 | 0.27 | 0.30 | 0.21 | 0.34 | 0.24 | 0.25 |
| TTC | 0.35 | 0.36 | 0.42 | 0.32 | 0.40 | 0.45 | 0.36 | 0.51 | 0.54 | 0.43 | 0.40 |
| TTT | 0.65 | 0.64 | 0.58 | 0.68 | 0.60 | 0.55 | 0.64 | 0.49 | 0.46 | 0.57 | 0.60 |
| CCA | 0.12 | 0.15 | 0.24 | 0.23 | 0.21 | 0.23 | 0.19 | 0.24 | 0.24 | 0.24 | 0.21 |
| CCC | 0.18 | 0.19 | 0.15 | 0.15 | 0.14 | 0.14 | 0.11 | 0.14 | 0.23 | 0.16 | 0.22 |
| CCG | 0.42 | 0.49 | 0.43 | 0.40 | 0.48 | 0.42 | 0.46 | 0.45 | 0.33 | 0.40 | 0.37 |
| CCT | 0.28 | 0.17 | 0.17 | 0.21 | 0.16 | 0.21 | 0.23 | 0.17 | 0.19 | 0.20 | 0.20 |
| AGC | 0.24 | 0.11 | 0.20 | 0.16 | 0.12 | 0.15 | 0.35 | 0.20 | 0.32 | 0.12 | 0.26 |
| AGT | 0.22 | 0.09 | 0.19 | 0.21 | 0.09 | 0.13 | 0.31 | 0.06 | 0.15 | 0.10 | 0.15 |
| TCA | 0.09 | 0.09 | 0.09 | 0.18 | 0.10 | 0.12 | 0.06 | 0.15 | 0.08 | 0.12 | 0.08 |
| TCC | 0.15 | 0.23 | 0.14 | 0.12 | 0.13 | 0.15 | 0.09 | 0.21 | 0.23 | 0.23 | 0.24 |
| TCG | 0.18 | 0.21 | 0.18 | 0.16 | 0.30 | 0.22 | 0.11 | 0.22 | 0.10 | 0.23 | 0.12 |
| TCT | 0.12 | 0.26 | 0.20 | 0.18 | 0.26 | 0.24 | 0.09 | 0.16 | 0.12 | 0.20 | 0.15 |
| ACA | 0.10 | 0.10 | 0.12 | 0.23 | 0.13 | 0.15 | 0.14 | 0.11 | 0.13 | 0.16 | 0.16 |
| ACC | 0.45 | 0.34 | 0.39 | 0.28 | 0.47 | 0.35 | 0.37 | 0.38 | 0.39 | 0.34 | 0.46 |
| ACG | 0.30 | 0.29 | 0.24 | 0.25 | 0.26 | 0.28 | 0.30 | 0.30 | 0.31 | 0.28 | 0.25 |
| ACT | 0.16 | 0.28 | 0.25 | 0.25 | 0.14 | 0.21 | 0.18 | 0.21 | 0.18 | 0.22 | 0.14 |
| TAC | 0.33 | 0.35 | 0.43 | 0.31 | 0.38 | 0.42 | 0.41 | 0.39 | 0.48 | 0.39 | 0.44 |
| TAT | 0.67 | 0.65 | 0.57 | 0.69 | 0.62 | 0.58 | 0.59 | 0.61 | 0.52 | 0.61 | 0.56 |
| GTA | 0.14 | 0.10 | 0.15 | 0.13 | 0.10 | 0.15 | 0.12 | 0.17 | 0.13 | 0.17 | 0.12 |
| GTC | 0.21 | 0.29 | 0.21 | 0.24 | 0.29 | 0.24 | 0.24 | 0.23 | 0.26 | 0.22 | 0.26 |
| GTG | 0.39 | 0.33 | 0.30 | 0.35 | 0.32 | 0.39 | 0.39 | 0.41 | 0.39 | 0.37 | 0.32 |
| GTT | 0.26 | 0.28 | 0.34 | 0.28 | 0.29 | 0.22 | 0.26 | 0.19 | 0.23 | 0.24 | 0.30 |

| **Concatenate** | **23** | **24** | **25** | **26** | **27** | **28** | **29** | **30** | **31** | **32** | **33** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (kb) | 308.05 | 321.78 | 341.66 | 310.31 | 333.66 | 370.78 | 326.68 | 324.28 | 329.25 | 339.14 | 268.67 |
| GC% | 0.53 | 0.53 | 0.54 | 0.53 | 0.53 | 0.53 | 0.53 | 0.53 | 0.53 | 0.53 | 0.52 |
| Identity% | 0.94 | 0.93 | 0.93 | 0.94 | 0.95 | 0.96 | 0.96 | 0.96 | 0.94 | 0.95 | 0.97 |
| Mean Fop | 0.46 | 0.47 | 0.48 | 0.49 | 0.50 | 0.51 | 0.53 | 0.54 | 0.57 | 0.60 | 0.68 |
| **Estimated Parameters** | | | | | | | | | | | |
| $\omega$ | 0.05 | 0.12 | 0.05 | 0.04 | 0.04 | 0.04 | 0.03 | 0.03 | 0.03 | 0.02 | 0.03 |
| $\pi_A$ | 0.36 | 0.32 | 0.33 | 0.33 | 0.26 | 0.31 | 0.33 | 0.28 | 0.26 | 0.28 | 0.29 |
| $\pi_C$ | 0.20 | 0.24 | 0.20 | 0.18 | 0.19 | 0.23 | 0.19 | 0.21 | 0.18 | 0.16 | 0.21 |
| $\pi_G$ | 0.19 | 0.21 | 0.19 | 0.20 | 0.23 | 0.20 | 0.18 | 0.22 | 0.21 | 0.19 | 0.15 |
| $\pi_T$ | 0.25 | 0.24 | 0.28 | 0.29 | 0.33 | 0.26 | 0.30 | 0.29 | 0.35 | 0.37 | 0.34 |
| Codon preferences $\phi_{aa}(i)$ | | | | | | | | | | | |
| GCA | 0.20 | 0.22 | 0.21 | 0.18 | 0.25 | 0.20 | 0.17 | 0.26 | 0.26 | 0.25 | 0.32 |
| GCC | 0.32 | 0.26 | 0.28 | 0.29 | 0.29 | 0.25 | 0.31 | 0.27 | 0.24 | 0.24 | 0.12 |
| GCG | 0.25 | 0.33 | 0.34 | 0.35 | 0.31 | 0.34 | 0.33 | 0.27 | 0.27 | 0.32 | 0.31 |
| GCT | 0.23 | 0.19 | 0.17 | 0.18 | 0.15 | 0.22 | 0.19 | 0.20 | 0.23 | 0.20 | 0.25 |
| AGA | 0.03 | 0.02 | 0.02 | 0.08 | 0.23 | 0.02 | 0.01 | 0.10 | 0.02 | 0.18 | 0.00 |
| AGG | 0.03 | 0.01 | 0.01 | 0.06 | 0.06 | 0.11 | 0.03 | 0.03 | 0.11 | 0.03 | 0.10 |
| CGA | 0.09 | 0.05 | 0.04 | 0.06 | 0.07 | 0.05 | 0.04 | 0.03 | 0.02 | 0.03 | 0.01 |
| CGC | 0.34 | 0.33 | 0.40 | 0.37 | 0.33 | 0.28 | 0.47 | 0.35 | 0.48 | 0.42 | 0.37 |
| CGG | 0.11 | 0.07 | 0.11 | 0.09 | 0.04 | 0.06 | 0.08 | 0.01 | 0.00 | 0.02 | 0.02 |
| CGT | 0.39 | 0.53 | 0.42 | 0.34 | 0.27 | 0.48 | 0.36 | 0.48 | 0.37 | 0.33 | 0.50 |
| AAC | 0.49 | 0.53 | 0.54 | 0.69 | 0.60 | 0.61 | 0.60 | 0.63 | 0.74 | 0.78 | 0.76 |
| AAT | 0.51 | 0.47 | 0.46 | 0.31 | 0.40 | 0.39 | 0.40 | 0.37 | 0.26 | 0.22 | 0.24 |
| GAC | 0.31 | 0.34 | 0.45 | 0.48 | 0.49 | 0.38 | 0.48 | 0.38 | 0.51 | 0.62 | 0.60 |
| GAT | 0.69 | 0.66 | 0.55 | 0.52 | 0.51 | 0.62 | 0.52 | 0.62 | 0.49 | 0.38 | 0.40 |
| TGC | 0.45 | 0.54 | 0.55 | 0.57 | 0.50 | 0.60 | 0.65 | 0.69 | 0.55 | 0.74 | 0.66 |
| TGT | 0.55 | 0.46 | 0.45 | 0.43 | 0.50 | 0.40 | 0.35 | 0.31 | 0.45 | 0.26 | 0.34 |
| CAA | 0.29 | 0.24 | 0.25 | 0.23 | 0.35 | 0.26 | 0.21 | 0.23 | 0.23 | 0.32 | 0.20 |
| CAG | 0.71 | 0.76 | 0.75 | 0.77 | 0.65 | 0.74 | 0.79 | 0.77 | 0.77 | 0.68 | 0.80 |
| GAA | 0.57 | 0.59 | 0.56 | 0.56 | 0.71 | 0.65 | 0.70 | 0.71 | 0.70 | 0.68 | 0.65 |
| GAG | 0.43 | 0.41 | 0.44 | 0.44 | 0.29 | 0.35 | 0.30 | 0.29 | 0.30 | 0.32 | 0.35 |
| GGA | 0.08 | 0.09 | 0.12 | 0.09 | 0.10 | 0.06 | 0.05 | 0.10 | 0.06 | 0.04 | 0.03 |
| GGC | 0.32 | 0.37 | 0.35 | 0.38 | 0.39 | 0.37 | 0.42 | 0.38 | 0.42 | 0.48 | 0.32 |
| GGG | 0.23 | 0.19 | 0.20 | 0.25 | 0.13 | 0.18 | 0.16 | 0.16 | 0.14 | 0.09 | 0.09 |
| GGT | 0.37 | 0.36 | 0.33 | 0.28 | 0.37 | 0.39 | 0.37 | 0.36 | 0.37 | 0.39 | 0.56 |
| CAC | 0.42 | 0.33 | 0.53 | 0.54 | 0.57 | 0.44 | 0.66 | 0.65 | 0.70 | 0.68 | 0.59 |
| CAT | 0.58 | 0.67 | 0.47 | 0.46 | 0.43 | 0.56 | 0.34 | 0.35 | 0.30 | 0.32 | 0.41 |
| ATA | 0.07 | 0.03 | 0.05 | 0.05 | 0.09 | 0.05 | 0.03 | 0.05 | 0.05 | 0.01 | 0.00 |
| ATC | 0.32 | 0.36 | 0.47 | 0.49 | 0.53 | 0.52 | 0.52 | 0.47 | 0.68 | 0.68 | 0.76 |
| ATT | 0.62 | 0.61 | 0.47 | 0.46 | 0.38 | 0.43 | 0.45 | 0.47 | 0.27 | 0.31 | 0.24 |
| CTA | 0.02 | 0.02 | 0.03 | 0.04 | 0.04 | 0.03 | 0.04 | 0.03 | 0.04 | 0.03 | 0.02 |
| CTC | 0.11 | 0.07 | 0.15 | 0.19 | 0.12 | 0.12 | 0.16 | 0.12 | 0.13 | 0.09 | 0.05 |
| CTG | 0.40 | 0.47 | 0.50 | 0.49 | 0.56 | 0.50 | 0.50 | 0.43 | 0.58 | 0.66 | 0.74 |
| CTT | 0.13 | 0.11 | 0.11 | 0.09 | 0.07 | 0.10 | 0.11 | 0.16 | 0.08 | 0.05 | 0.05 |
| TTA | 0.11 | 0.15 | 0.08 | 0.07 | 0.10 | 0.09 | 0.08 | 0.13 | 0.06 | 0.07 | 0.04 |
| TTG | 0.22 | 0.18 | 0.14 | 0.12 | 0.11 | 0.16 | 0.12 | 0.13 | 0.10 | 0.10 | 0.10 |
| AAA | 0.70 | 0.61 | 0.68 | 0.78 | 0.81 | 0.69 | 0.69 | 0.76 | 0.82 | 0.73 | 0.67 |
| AAG | 0.30 | 0.39 | 0.32 | 0.22 | 0.19 | 0.31 | 0.31 | 0.24 | 0.18 | 0.27 | 0.33 |
| TTC | 0.41 | 0.38 | 0.54 | 0.52 | 0.56 | 0.55 | 0.56 | 0.60 | 0.75 | 0.77 | 0.73 |
| TTT | 0.59 | 0.62 | 0.46 | 0.48 | 0.44 | 0.45 | 0.44 | 0.40 | 0.25 | 0.23 | 0.27 |
| CCA | 0.17 | 0.25 | 0.18 | 0.22 | 0.28 | 0.30 | 0.16 | 0.27 | 0.31 | 0.34 | 0.27 |
| CCC | 0.11 | 0.10 | 0.14 | 0.13 | 0.13 | 0.07 | 0.13 | 0.05 | 0.06 | 0.03 | 0.01 |
| CCG | 0.36 | 0.48 | 0.38 | 0.46 | 0.45 | 0.49 | 0.50 | 0.59 | 0.56 | 0.47 | 0.66 |
| CCT | 0.36 | 0.16 | 0.29 | 0.19 | 0.14 | 0.14 | 0.21 | 0.08 | 0.07 | 0.16 | 0.06 |
| AGC | 0.07 | 0.15 | 0.16 | 0.17 | 0.32 | 0.09 | 0.33 | 0.12 | 0.18 | 0.19 | 0.01 |
| AGT | 0.07 | 0.10 | 0.09 | 0.08 | 0.14 | 0.07 | 0.15 | 0.05 | 0.10 | 0.04 | 0.00 |
| TCA | 0.12 | 0.11 | 0.12 | 0.14 | 0.08 | 0.24 | 0.08 | 0.14 | 0.10 | 0.10 | 0.05 |
| TCC | 0.31 | 0.16 | 0.27 | 0.21 | 0.21 | 0.17 | 0.20 | 0.24 | 0.32 | 0.31 | 0.35 |
| TCG | 0.20 | 0.22 | 0.19 | 0.23 | 0.15 | 0.22 | 0.12 | 0.14 | 0.11 | 0.05 | 0.14 |
| TCT | 0.24 | 0.25 | 0.17 | 0.17 | 0.11 | 0.21 | 0.12 | 0.30 | 0.20 | 0.31 | 0.46 |
| ACA | 0.15 | 0.14 | 0.14 | 0.10 | 0.14 | 0.10 | 0.11 | 0.13 | 0.09 | 0.09 | 0.13 |
| ACC | 0.46 | 0.40 | 0.36 | 0.36 | 0.49 | 0.42 | 0.41 | 0.39 | 0.54 | 0.45 | 0.42 |
| ACG | 0.16 | 0.25 | 0.30 | 0.29 | 0.19 | 0.26 | 0.27 | 0.15 | 0.17 | 0.24 | 0.16 |
| ACT | 0.23 | 0.21 | 0.21 | 0.25 | 0.17 | 0.22 | 0.21 | 0.32 | 0.20 | 0.22 | 0.29 |
| TAC | 0.55 | 0.35 | 0.48 | 0.62 | 0.60 | 0.46 | 0.50 | 0.48 | 0.50 | 0.69 | 0.57 |
| TAT | 0.45 | 0.65 | 0.52 | 0.38 | 0.40 | 0.54 | 0.50 | 0.52 | 0.50 | 0.31 | 0.43 |
| GTA | 0.13 | 0.15 | 0.16 | 0.13 | 0.24 | 0.16 | 0.12 | 0.22 | 0.18 | 0.20 | 0.27 |
| GTC | 0.26 | 0.17 | 0.21 | 0.31 | 0.23 | 0.17 | 0.27 | 0.19 | 0.23 | 0.22 | 0.10 |
| GTG | 0.35 | 0.40 | 0.44 | 0.39 | 0.33 | 0.42 | 0.31 | 0.31 | 0.34 | 0.31 | 0.32 |
| GTT | 0.25 | 0.27 | 0.19 | 0.17 | 0.21 | 0.25 | 0.29 | 0.27 | 0.25 | 0.27 | 0.31 |

**Table 1.** Table showing descriptive statistics of the gene concatenates along with the value of estimated parameters of the model. Identity% is the fraction of exactly identical positions in the alignments. Codons are organized by synonymous groups for easier reading.

# A phylogenomic test of the hypotheses for the origin of eukaryotes

Nicolas C. ROCHETTE[1], Céline BROCHIER-ARMANET[1] and Manolo GOUY[1]

Laboratoire de Biométrie et Biologie Évolutive,
UMR5558 CNRS, Université Claude Bernard, 43 bd du 11 novembre 1918, 69622 VILLEURBANNE cedex, France
nicolas.rochette|celine.brochier-armanet|manolo.gouy@univ-lyon1.fr

**Abstract** *The evolutionary origin of eukaryotes is a question of great interest for which several widely different hypotheses have been proposed. These hypotheses predict distinct patterns of evolutionary relationships for individual genes of the ancestral eukaryotic genome. The availability of numerous completely sequenced genomes covering the three domains of life allows to challenge these predictions.*

*We performed a systematic study of the phylogenetic relationships between ancestral eukaryotic genes and Archaea and Bacteria, with high methodological standards. Moreover, we developed novel principles and methods in order to account for the constant remodelling of prokaryotic genomes by horizontal gene transfer and gene loss, what, we show, is crucial to the correct interpretation of gene trees spanning such a large evolutionary scale.*

*Our analysis very clearly recovered the two established properties of the eukaryotic genome : its apparent bacterial-archaeal mosaicism, and the alphaproteobacterial origin of mitochondria. We then demonstrate that (i) the set of genes that link to alphaproteobacteria consists almost exclusively of genes involved in mitochondrial respiration and protein processing, and (ii) there is no evidence for a relation of eukaryotes with a specific bacterial lineage other than alphaproteobacteria. Moreover, our results suggest that eukaryotes may branch among Archaea, though we exclude a close association with Euryarchaeota or Crenarchaeota.*

*Overall, our observations are compatible only with some of the "early-mitochondria" hypotheses, which involve an early endosymbiosis of an alphaproteobacterium in an archaeal host, and with a few of the "autogenous" hypotheses, notably the slow-drip hypothesis in which proto-eukaryotes were, like prokaryotes, very prone to horizontal gene transfers.*

**Keywords** evolution ; phylogenomics ; horizontal gene transfer ; genome-content variability; archaea

## 1 Introduction

All known cellular organisms belong to one of three domains : Bacteria, Archaea or Eukarya. These three groups share a common ancestry, but each also harbor distinctive features whose origin and evolution are not well understood. The origin of eukaryotes, in particular, has drawn much attention. Eukaryotes have an elaborate cell biology, and no evolutionary intermediates are known : rather, the last eukaryotic common ancestor (LECA) is though to have had essentially all the emblematic eukaryotic characters, including a nucleus [1], actin and tubulin [2, 3], mitochondria [4], and the sexual cycle [5]. A great diversity of hypotheses for the origin of eukaryotes have been proposed. The objective of this study is to test them using the distinctive predictions they make regarding the lineages that were involved in early eukaryotic evolution, and to which eukaryotic genes should relate.

For instance, one famous hypothesis, the "syntrophy hypothesis" [6], proposes that Eukarya are a chimera of a euryarchaeon and a deltaproteobacterium, with an alphaproteobacterial endosymbiont. Therefore it predicts that eukaryotic genes, when they have prokaryotic homologs, should be mainly related to euryarchaeal, deltaproteobacterial and alphaproteobacterial genes (through endosymbiotic gene transfers (EGTs)). Similarly, under the very famous "hydrogen hypothesis" [7], eukaryotic genes are expected to have been inherited from the alphaproteobacterial ancestor of mitochondria and from the methanogenic euryarchaeon which hosted it.

Cavalier-Smith's Neomura hypothesis [8] assumes that Eukarya are the sister group of all Archaea and explains the existence of (apparently) bacteria-related genes in Eukarya by EGTs and losses of LUCA genes in the ancestors of Archaea, what results in Eukarya and Bacteria both having a gene Archaea lack. Finally, some hypotheses propose that the eukaryotic lineage acquired many bacterial genes by horizontal gene transfer (HGT), as for instance the slow-drip hypothesis [9] which proposes that ancient eukaryotic ancestors often acquired new genes through HGT, like prokaryotes nowadays.

Investigating the phylogenetic relationships between eukaryotic and prokaryotic genes on a genomic scale is thus an essential piece in the debate over the origin of Eukarya. This question has been addressed many times in the past fifteen years [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20] but remains essentially unresolved.

We undertook to dissect the origins of eukaryotic genes at a much finer scale than has been performed in previous studies. In particular, we could introduce a distinction between genes which phylogeny really supports a relationship between eukaryotes and a particular prokaryotic taxonomic group, and genes whose evolutionary histories are blurred by HGTs among prokaryotes or have poorly resolved phylogenies. For this we formalized interpretation principles that are of interest for all phylogenetic studies facing extensive HGT and genome-content variability issues.

## 2 Results and discussion

### 2.1 Identification of LECA clades, taxonomic sampling and phylogenetic inferences

We started our work from the Hogenom (v5) database, which provides clusters of homologous sequences built from 946 complete genomes from the three domains of life [21], as well as maximum-likelihood trees of those clusters. We retrieved the clusters of homologs that contained sequences of both eukaryotes and prokaryotes. We then identified 554 clades of monophyletic eukaryotic sequences traceable to LECA (see Methods). Each of those "LECA clades" corresponds to (at least) one gene in the genome of LECA.

The initial clusters were very large and taxonomically unbalanced (as they include all sequenced genomes). Also, we restricted our analysis to 144 and 39 representative genomes of Bacteria and Archaea, respectively. For each LECA clade and its prokaryotic homologs, a maximum-likelihood tree and 100 bootstrap trees were computed.

### 2.2 Analysis through "configurations"

The trees were extremely heterogeneous in terms of species presence or absence, number of paralogs in individual genomes, branching patterns between related genomes, as well as in terms of topology, branch length, and bootstrap support distribution among branches. This extensive diversity rendered the definition of a generic strategy very challenging. Previous studies relied either on the best-BLAST-hit criterion [11, 12, 13] or on a crude sister-group criterion [17, 18, 20]. In contrast, we found that a proper interpretation of the underlying gene trees required to consider the taxonomic context of the branching point of eukaryotes among prokaryotes and to control the overall taxonomic sampling.

Therefore, our conclusions were based on non-trivial topological criteria we named "configurations". In a word, configurations are designed to discriminate unambiguous scenarios such as the one presented in Fig. 1 from scenarios that were likely confused by recent horizontal gene transfers. This is accomplished by considering the the taxonomic representativeness of the putative prokaryotic relatives of a LECA clade.

### 2.3 Archaeal-bacterial mosaicism

Using "configurations", we annotated every bootstrap tree for each of the LECA clades. Our LECA genome appeared partly archaea-related and partly bacteria-related, with proportions one third and two thirds respectively (Fig. 2), in agreement with previous reports [11, 19, 20]. We also recovered the functional contrast between archaea-related and bacteria-related LECA clades [22] : an overwhelming majority of the metabolic LECA clades were bacteria-related, while genes involved in replication, transcription and translation were archaea-related (except for those functioning in the mitochondria). Other processes were either of mixed origin (e.g. DNA repair) or of unclear origin (e.g. membrane transport).

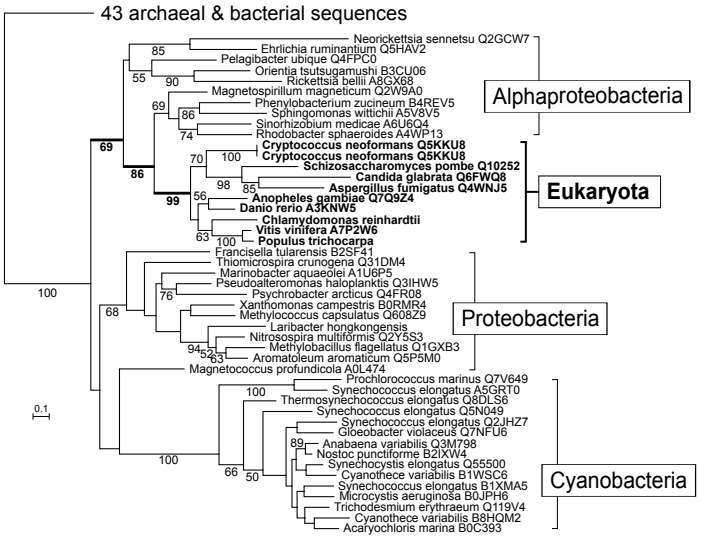



**Figure 1.** The maximum likelihood tree for the hydroxybenzoate polyprenyltransferase (COQ2) LECA clade matches the "alphaproteobacteria" configuration.

## 2.4 Relations of Eukarya to Archaea

Strikingly, only 4 LECA clades out of 120 archaea-related ones supported a "three-domains" topology (Fig. 2). However, because the phylogenetic reconstructions for archaea-related LECA clades often suffered lack of signal, this result does not actually exclude the three-domains hypothesis for the tree of life. Nevertheless, it clearly does not favor it, and is clearly compatible with a within-archaea branching of Eukarya.

Furthermore, there was essentially no evidence for a close relationship between Eukarya and Crenarchaeota or Euryarchaeota (Fig. 2). This was not a consequence of basic limitations of the method, as it was able, for example, to unambiguously identify the basal Thermococcales or the relatively fast-evolving Thermoplasma as members of Euryarchaeota (not shown). In contrast, the pattern recovered for the phylum Thaumarchaeota (not shown) was similar to the one observed for Eukarya.

These results question that eukaryotes could derive from a methanogen, as some hypotheses suggest [6, 7]. Indeed, methanogenesis is though to have evolved only in Euryarchaeota [23], to which eukaryotes appear not to be related in any way.

## 2.5 Relations of eukaryotes to bacterial phyla

As expected, given that the mitochondrion is a derived alphaproteobacterium, a substantial number of LECA clades (38) were found to be associated with clades of alphaproteobacterial sequences. Interestingly, almost all of them (35) were involved in core mitochondrial functions such as respiration and protein processing.

In addition, our analysis identified a few clades that seemed to be related to bacterial phyla other than alphaproteobacteria. This scenario was supported for six LECA clades, which were related to respectively Cyanobacteria (4), Chlamydiae (1) and Verrucomicrobiae (1).

These figures are modest in regard of the whole 242 bacteria-related LECA clades. Most (198) of these clades, although they are clearly related to Bacteria, were eventually not reliably traceable to a particular phylum. These cases were referred to as "bacterial-domain-related". In a general manner, they can be explained by either lack of phylogenetic signal or HGTs among prokaryotes (or a combination of both). We investigated the role of lack of signal and found that it could be excluded in many cases. Thus HGTs were likely the primary cause for the commonness of "bacterial-domain-related" among LECA-clade annotations.

These results —a few alphaproteobacteria-related genes and many genes with complex histories— contrast sharply with earlier studies [11, 13, 14, 17, 18, 20], where eukaryotic genes appeared related to all sorts of Bacteria (including Alphaproteobacteria). And remarkably, if we disregarded "configurations" and opted for

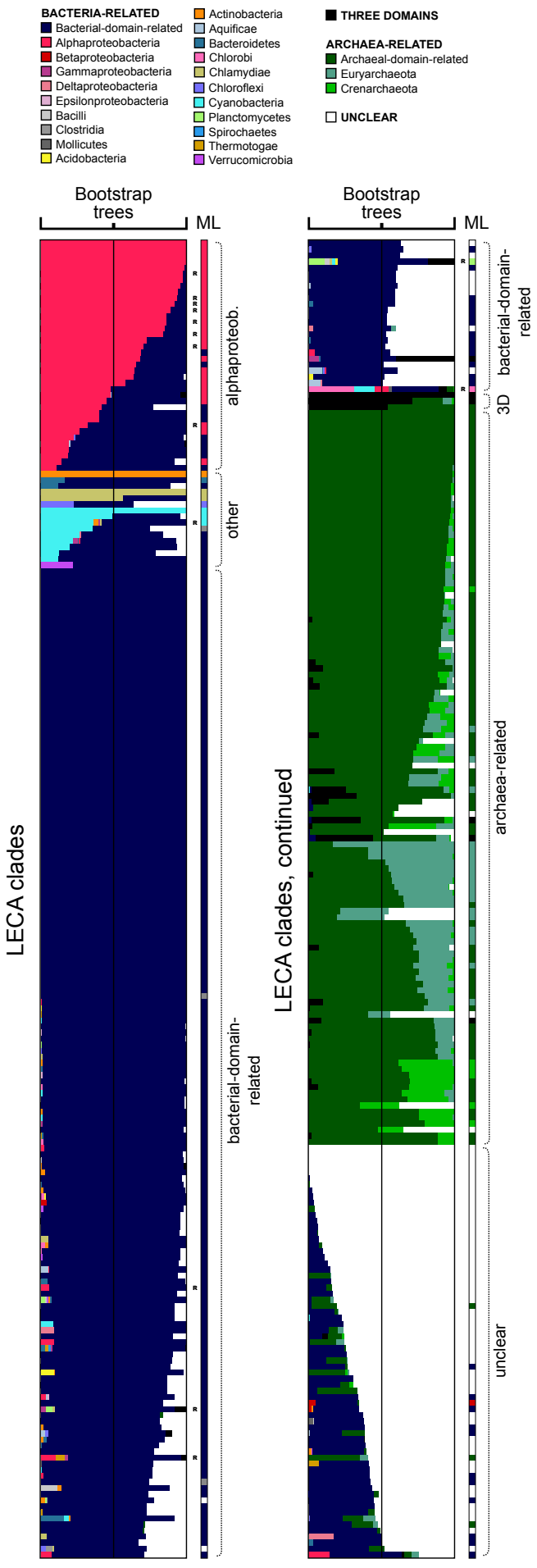



**Figure 2.** The origins of eukaryotic genes. Each color represents a configuration as indicated, and each row corresponds to a LECA clade. The configurations of all ML trees are given in the vertical bar on the right, while the widths of the color bars in the wider block correspond to the support for the configuration among bootstrap trees (i.e. the proportion of trees in which it appears). LECA clades are sorted by configuration and decreasing support. Overall, 38 LECA clades were traceable to alphaproteobacteria, 6 to diverse bacterial phyla, 197 to Bacteria though not to a particular phylum or class ("bacterial-domain-related"), 3 appeared in the "three domains" configuration, 117 linked to archaea, and 71 could not be linked to a particular domain ("unclear").

the naive sister-group-identity criterion used in those studies, we observed a pattern similar to theirs, which thus appears to be spurious.

Hence one major result brought about by our study is that there is no phylogenetic evidence that a particular bacterial lineage (apart from Alphaproteobacteria) was involved in the origin of eukaryotes. This observation is remarkable as many hypotheses advocate that bacteria-related eukaryotic genes descend in part from the ancestor of mitochondria, and in part from (an)other bacterial lineage(s). These scenarios thus receive no support. Moreover, though additional LECA clades of prokaryotic origin could certainly be identified in the future, it is unlikely that a large-scale signal would have been missed.

## 3 Conclusion

The mosaicism of the eukaryotic genome is challenging. We demonstrate why determining the histories of these genes precisely is difficult, and often impossible by current means of analysis. Nevertheless, our analysis establishes that there is no phylogenomic support in favor of "fusion" hypotheses. In addition, we present evidence that single-gene phylogenies collectively exclude a close relationship between Eukarya and Crenarchaeota or Euryarchaeota, and suggest that Eukarya branch in a basal position within Archaea. Finally, we show that the slow-drip and early-mitochondria hypotheses are compatible with current genomic data under certain assumptions.

Further progress on the question of the origin of eukaryotes is expected to arise from new genome sequences in undersampled archaeal and eukaryotic phyla, improved methods for reconstructing taxon-rich single-gene phylogenies (matrix mixture models [24] are a step in this direction) and better knowledge of the biology of Bacteria and Archaea [25, 26, 27, 28] and of that of the last eukaryotic common ancestor [2, 29, 30].

## 4 Methods

### 4.1 Identification of LECA clades

The Hogenom (v5) database includes all proteins from 64 eukaryotic, 62 archaeal and 820 bacterial complete genomes, and provides pre-computed clusters of homologs based on all-vs-all BLASTs and transitive homology bonds [21, 31], and ML trees of all clusters. Clusters containing at least two of Opisthokonts, Plantae and Chromalveolates, and one prokaryotic phylum were retrieved. All clades of monophyletic eukaryotic sequences were extracted by means of custom tree-parsing algorithms. Eukaryotic clades were then inferred to trace back to LECA if they contained sequences from at least (i) two unikont species and two Plantae, (ii) two Unikonts and two Chromalveolates, or (iii) two Plantae, two Chromalveolates and one kinetoplastid. Because recent eukaryotes-to-prokaryotes HGTs may confuse this strategy by making eukaryotes appear paraphyletic, all trees were manually inspected before eukaryotic clades were extracted, and isolated prokaryotic sequences branching within a group of diverse eukaryotes were removed.

### 4.2 Selection of representative archaeal and bacterial genomes

All analyses except the identification of LECA clades were performed using the same subset of 183 representative archaeal and bacterial genomes. In Archaea, one genome was sampled in each represented genus, except *Nanoarchaeum equitans* which was not included because of its high evolutionary rate and uncertain phylogenetic position, for a total of 39 genomes. In Bacteria, up to 15 genomes were sampled for each phylum according to a reference species phylogeny [32], except for Proteobacteria and Firmicutes which were sampled class-wise. For bacterial phyla for which genomes were available for less than 15 genera, one genome was randomly sampled in each genus. Overall, 144 bacterial genomes were included.

### 4.3 Phylogenetic inferences

The results presented in the figure were obtained using Probcons (default parameters) [33], BMGE (BLOSSUM30 matrix) [34], and RAxML (CAT rates, LG matrix, 100 nonparametric bootstrap replicates) [35].

### 4.4 Configurations

The "configuration" of every bootstrap or ML tree was determined as follows. A LECA clade was said to be related to a particular phylum (or class for Proteobacteria and Firmicutes) if (i) it branched inside a clade of monophyletic sequences of this phylum, and (ii) that these sequences represented more than half the species of the group. Similarly, a LECA clade was said to be bacteria-related (respectively archaea-related) if it branched inside a clade of bacterial (respectively archaeal) sequences representing at least 10 species. When a LECA clade was bacteria-related (respectively archaea-related) but could not be related to any phylum, it was said to be just "bacterial-domain-related" (respectively "archaeal-domain-related"). A tree was labeled "three-domains" if all three domains were monophyletic and at least 10 archaeal and 10 bacterial species were represented. A tree in which the LECA clade was neither bacteria-related, nor archaea-related, nor in a three-domains position, was labeled "unclear". Trees in which the representative sequences for a LECA clade were paraphyletic were labeled "paraphyletic" and discarded. The identification of configurations was implemented using the Bio++ [36] C++ library. Source code is available upon request.

### 4.5 Inspection of LECA clades putatively related to bacterial groups other than alphaproteobacteria

The cases of these clades were investigated in more details. First, the taxonomic distribution of sequences in the 183-prokaryotes tree was compared to the one of the full 882-prokaryotes cluster, in order to check the representativeness of the smaller genome sample. In addition, the reliability of the HOGENOM clustering was checked by performing a HMMER 3.0 (http://hmmer.org) search in 183 complete proteomes, using as seed a MAFFT (default FFT-NS-2 mode) alignment of the 183-prokaryotes cluster. Finally, we reviewed the maximum likelihood tree, considering the taxonomic distribution, potential HGTs, and bootstrap support values.

## References


[1] B. J. Mans, V. Anantharaman, L. Aravind, and E. V. Koonin, "Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex," *Cell cycle (Georgetown, Tex.)*, vol. 3, pp. 1612–1637, Dec. 2004. PMID: 15611647.

[2] N. Yutin, M. Y. Wolf, Y. I. Wolf, and E. V. Koonin, "The origins of phagocytosis and eukaryogenesis," *Biology Direct*, vol. 4, no. 1, p. 9, 2009.

[3] B. Hammesfahr and M. Kollmar, "Evolution of the eukaryotic dynactin complex, the activator of cytoplasmic dynein," *BMC Evolutionary Biology*, vol. 12, p. 95, June 2012.

[4] T. Gabaldón and M. A. Huynen, "From endosymbiont to host-controlled organelle: The hijacking of mitochondrial protein synthesis and metabolism," *PLoS Computational Biology*, vol. 3, no. 11, p. e219, 2007.

[5] M. A. Ramesh, S.-B. Malik, and J. Logsdon, John M, "A phylogenomic inventory of meiotic genes; evidence for sex in giardia and an early eukaryotic origin of meiosis," *Current biology: CB*, vol. 15, pp. 185–191, Jan. 2005. PMID: 15668177.

[6] P. Lopez-Garcia and D. Moreira, "Selective forces for the origin of the eukaryotic nucleus," *BioEssays*, vol. 28, pp. 525–533, May 2006.

[7] W. Martin and M. Müller, "The hydrogen hypothesis for the first eukaryote," *Nature*, vol. 392, pp. 37–41, Mar. 1998.

[8] T. Cavalier-Smith, "Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution," *Biology direct*, vol. 5, p. 7, 2010. PMID: 20132544.

[9] L. Lester, A. Meade, and M. Pagel, "The slow road to the eukaryotic genome," *BioEssays: news and reviews in molecular, cellular and developmental biology*, vol. 28, pp. 57–64, Jan. 2006. PMID: 16369937.

[10] T. Horiike, K. Hamada, S. Kanaya, and T. Shinozawa, "Origin of eukaryotic cell nuclei by symbiosis of archaea in bacteria is revealed by homology-hit analysis," *Nature Cell Biology*, vol. 3, pp. 210–214, Feb. 2001. PMID: 11175755.

[11] C. Esser, N. Ahmadinejad, C. Wiegand, C. Rotte, F. Sebastiani, G. Gelius-Dietrich, K. Henze, E. Kretschmann, E. Richly, D. Leister, D. Bryant, M. A. Steel, P. J. Lockhart, D. Penny, and W. Martin, "A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes," *Molecular Biology and Evolution*, vol. 21, pp. 1643–1660, Sept. 2004. PMID: 15155797.

[12] A. Atteia, A. Adrait, S. Brugiere, M. Tardif, R. van Lis, O. Deusch, T. Dagan, L. Kuhn, B. Gontero, W. Martin, J. Garin, J. Joyard, and N. Rolland, "A proteomic survey of chlamydomonas reinhardtii mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the $\alpha$-proteobacterial mitochondrial ancestor," *Molecular Biology and Evolution*, vol. 26, pp. 1533–1548, Apr. 2009.

[13] E. V. Koonin, "The origin and early evolution of eukaryotes in the light of phylogenomics," *Genome Biology*, vol. 11, no. 5, p. 209, 2010. PMID: 20441612.

[14] R. Szklarczyk and M. A. Huynen, "Mosaic origin of the mitochondrial proteome," *PROTEOMICS*, vol. 10, pp. 4012–4024, Nov. 2010.

[15] M. C. Rivera and J. A. Lake, "The ring of life provides evidence for a genome fusion origin of eukaryotes," *Nature*, vol. 431, pp. 152–155, Sept. 2004. PMID: 15356622.

[16] O. Zhaxybayeva, L. Hamel, J. Raymond, and J. P. Gogarten, "Visualization of the phylogenetic content of five genomes using dekapentagonal maps," *Genome biology*, vol. 5, no. 3, p. R20, 2004. PMID: 15003123.

[17] D. Pisani, J. A. Cotton, and J. O. McInerney, "Supertrees disentangle the chimerical origin of eukaryotic genomes," *Molecular Biology and Evolution*, vol. 24, pp. 1752–1760, Apr. 2007.

[18] S. Saruhashi, K. Hamada, D. Miyata, T. Horiike, and T. Shinozawa, "Comprehensive analysis of the origin of eukaryotic genomes," *Genes & Genetic Systems*, vol. 83, pp. 285–291, Aug. 2008. PMID: 18931454.

[19] N. Yutin, K. S. Makarova, S. L. Mekhedov, Y. I. Wolf, and E. V. Koonin, "The deep archaeal roots of eukaryotes," *Molecular Biology and Evolution*, vol. 25, pp. 1619–1630, Apr. 2008.

[20] T. Thiergart, G. Landan, M. Schenk, T. Dagan, and W. F. Martin, "An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin," *Genome Biology and Evolution*, vol. 4, pp. 466–485, Feb. 2012.

[21] S. Penel, A.-M. Arigon, J.-F. Dufayard, A.-S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perrière, "Databases of homologous gene families for comparative genomics," *BMC Bioinformatics*, vol. 10, no. Suppl 6, p. S3, 2009.

[22] M. C. Rivera, R. Jain, J. E. Moore, and J. A. Lake, "Genomic evidence for two functionally distinct gene classes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, pp. 6239–6244, May 1998. PMID: 9600949.

[23] E. Bapteste, C. Brochier, and Y. Boucher, "Higher-level classification of the archaea: evolution of methanogenesis and methanogens," *Archaea (Vancouver, B.C.)*, vol. 1, pp. 353–363, May 2005. PMID: 15876569.

[24] S. Q. Le, N. Lartillot, and O. Gascuel, "Phylogenetic mixture models for proteins," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 363, pp. 3965–3976, Dec. 2008. PMID: 18852096.

[25] M. Lindsay, R. Webb, M. Strous, M. Jetten, M. Butler, R. Forde, and J. Fuerst, "Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell," *Archives of Microbiology*, vol. 175, pp. 413–429, June 2001.

[26] D. Hasenohrl, R. Konrat, and U. Blasi, "Identification of an RNase j ortholog in sulfolobus solfataricus: Implications for 5'-to-3' directional decay and 5'-end protection of mRNA in crenarchaeota," *RNA*, vol. 17, pp. 99–107, Nov. 2010.

[27] H. Shimada and A. Yamagishi, "Stability of heterochiral hybrid membrane made of bacterial sn -G3P lipids and archaeal sn -G1P lipids," *Biochemistry*, vol. 50, pp. 4114–4120, May 2011.

[28] N. Yutin and E. V. Koonin, "Archaeal origin of tubulin," *Biology Direct*, vol. 7, p. 10, Mar. 2012.

[29] E. Bapteste, R. L. Charlebois, D. MacLeod, and C. Brochier, "The two tempos of nuclear pore complex evolution: highly adapting proteins in an ancient frozen structure," *Genome biology*, vol. 6, no. 10, p. R85, 2005. PMID: 16207356.

[30] J. B. Dacks, A. A. Peden, and M. C. Field, "Evolution of specificity in the eukaryotic endomembrane system," *The International Journal of Biochemistry & Cell Biology*, vol. 41, pp. 330–340, Feb. 2009.

[31] V. Miele, S. Penel, V. Daubin, F. Picard, D. Kahn, and L. Duret, "High-quality sequence clustering guided by network topology and multiple alignment likelihood," *Bioinformatics*, vol. 28, pp. 1078–1085, Feb. 2012.

[32] D. Wu, P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J.-F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H.-P. Klenk, and J. A. Eisen, "A phylogeny-driven genomic encyclopaedia of bacteria and archaea," *Nature*, vol. 462, pp. 1056–1060, Dec. 2009.

[33] C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou, "ProbCons: probabilistic consistency-based multiple sequence alignment," *Genome research*, vol. 15, pp. 330–340, Feb. 2005. PMID: 15687296.

[34] A. Criscuolo and S. Gribaldo, "BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments," *BMC Evolutionary Biology*, vol. 10, no. 1, p. 210, 2010.

[35] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, pp. 2688–2690, Aug. 2006.

[36] J. Dutheil, S. Gaillard, E. Bazin, S. Glémin, V. Ranwez, N. Galtier, and K. Belkhir, "Bio++: a set of c++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics," *BMC bioinformatics*, vol. 7, p. 188, 2006. PMID: 16594991.

# Session 6C : Réseaux biologiques

# Properties of Random Complex Chemical Reaction Networks and Their Relevance to Biological Toy Models

Erwan BIGAN[1, 2], Jean-Marc STEYAERT[1] and Stéphane DOUADY[2]

[1] Laboratoire d'Informatique (LIX), École Polytechnique, 91128 Palaiseau Cedex, France

[2] Laboratoire Matière & Systèmes Complexes, UMR7057 CNRS, Université Paris Diderot, 75205 Paris Cedex 13, France

erwan.bigan@m4x.org

**Abstract** *We investigate the properties of large random conservative chemical reaction networks composed of elementary reactions endowed with either mass-action or saturating kinetics, assigning kinetic parameters in a thermodynamically-consistent manner. We find that such complex networks exhibit qualitatively similar behavior when fed with external nutrient flux. The nutrient is preferentially transformed into one specific chemical that is an intrinsic property of the network. We propose a self-consistent proto-cell toy model in which the preferentially synthesized chemical is a precursor for the cell membrane, and show that such proto-cells can exhibit sustainable homeostatic growth when fed with any nutrient diffusing through the membrane, provided that nutrient is metabolized at a sufficient rate.*

**Keywords** Chemical reaction network, biological toy model, homeostasis.

## 1 Introduction

Toy models help understand minimal requirements of life. Significant research has already been devoted to this field using random chemical reaction networks [1, 2, 3]. We build upon this previous work, extending this approach to random mass conservative chemical reaction networks with arbitrary stoichiometry.

Considering an arbitrary set of chemicals with randomly assigned standard Gibbs free energies of formation, random conservative networks of elementary reactions are generated, and kinetic parameters are randomly assigned to direct reactions, while kinetic parameters for reverse reactions are derived in a thermodynamically consistent manner. In addition to conventional mass-action kinetics, we also consider saturating kinetics. This is because mass-action kinetics only hold for diluted media, whereas biological systems are typically dense and macromolecular crowding is known to reduce effective reaction rates in the diffusion-limited regime [4].

This paper is organized as follows: in section 2, we describe our random reaction network model; in section 3, we characterize generated reaction networks in terms of equilibrium concentrations versus total mass density (section 3.1) and in terms of behavior under external nutrient flux (section 3.2); in section 4, we describe a proto-cell model enclosing such random reaction networks within a membrane (section 4.1) and explore the range of parameters within which cellular growth can be sustained (section 4.2); section 5 is devoted to discussion and conclusion.

## 2 Random Conservative Reaction Network Model

N different chemicals $\{A_i\}_{i=0,\ \ldots,\ N-1}$ are present in the system. Reactions are decomposed in monomolecular or bimolecular elementary reactions:

$$A_i + A_j \rightarrow A_k + A_l \text{ (exchange)}$$

$$A_i + A_j \rightarrow A_k \text{ (association)}$$

$$A_i \rightarrow A_k + A_l \text{ (dissociation)}$$

$$A_i \rightarrow A_k \text{ (transformation)}$$

It can be easily shown that any reactions with higher-order stoichiometry can be decomposed in such elementary reactions (at the expense on increasing the total number of reactions, and of increasing the total number of chemicals – typically intermediate composite compounds). Reactions with i=j or i=k are allowed, the only constraint is that the same expression should not be found on both sides.

All results presented in this paper have been obtained (i) excluding exchange reactions from the set, considering that they can be represented as successive combinations of association and dissociation reactions; and (ii) using N=10 as total number of chemicals (similar qualitative behavior has been observed with N=20, at the expense of longer computational time).

## 2.1 Topology

We use a previously published standard algorithm to check whether a network is conservative or not [5]. Running this algorithm also delivers the generating vectors of the convex cone of admissible mass vectors (a vector m with components $\{m_i\}_{i=0, \ldots, N-1}$ is an admissible mass vector if and only if $m^T S=0$ where $^T$ denotes the transpose operation and S is the N x R stoichiometry matrix, with R being the total number of reactions). To generate a random chemical reaction network, we proceed as follows: we choose a first random reaction, and then successively add new randomly chosen reactions while checking at each step that the network is conservative.

We observe that once the number of reactions exceeds N by a handful, the number of generating vectors of the convex cone falls to one, i.e. up to a multiplying factor there is only one single admissible mass vector for the given reaction network. At significantly larger number of reactions, we also observe that conservative networks reach a maximum size above which none of the remaining elementary reactions is orthogonal to the single mass vector, and that this maximum size is variable (ranging from 17 to 96 with an average of 42 direct reactions for a sample of 100 networks). Figure 1 shows an example of such a maximum-sized random conservative chemical reaction network that will be used as reference example in the remainder of this paper.

## 2.2 Kinetics

Standard Gibbs free energies of formation $\{G_i\}_{i=0, \ldots, N-1}$ are assigned to chemicals $\{A_i\}_{i=0, \ldots, N-1}$. For a given possible reaction of the constructed network, the forward versus reverse directions are determined by comparing the sum of standard Gibbs free energies of formation for reactants versus products.

Mass-action kinetics are characterized by a kinetic coefficient $k_r$ such that the reaction rate $f_r$ is given by $f_r=k_r[A_i][A_j]$ (resp. $f_r=k_r[A_i]$) for a bimolecular (resp. monomolecular) reaction. $k_r^{\rightarrow}$ are first assigned to direct reactions, and $k_r^{\leftarrow}$ for reverse reactions are then derived as given in the expressions below, where $\Delta G$ denotes the change in Gibbs free energy for the considered reaction, R the ideal gas constant, and c° the standard concentration:

- Monomolecular forward ($A_i \rightarrow \ldots$): $k_r^{\rightarrow}=k_{avg\_monomolecular}10^{random(-s/2, s/2)}$
  - Monomolecular reverse ($A_i \leftarrow A_k$): $k_r^{\leftarrow}=k_r^{\rightarrow}\exp(-|\Delta G|/RT)$
  - Bimolecular reverse ($A_i \leftarrow A_k+A_l$): $k_r^{\leftarrow}=(k_r^{\rightarrow}/c°)\exp(-|\Delta G|/RT)$
- Bimolecular forward ($A_i+A_j \rightarrow \ldots$): $k_r^{\rightarrow}=k_{avg\_bimolecular}10^{random(-s/2, s/2)}$
  - Monomolecular reverse ($A_i+A_j \leftarrow A_k$): $k_r^{\leftarrow}=c°k_r^{\rightarrow}\exp(-|\Delta G|/RT)$
  - Bimolecular reverse ($A_i+A_j \leftarrow A_k+A_l$): $k_r^{\leftarrow}=k_r^{\rightarrow}\exp(-|\Delta G|/RT)$

with random(-s/2, s/2) designating a random real number between -s/2 and s/2 where s is the spread of kinetic parameters for direct reactions, counted in orders of magnitude.

Saturating kinetics are simply derived from mass-action kinetics by introducing a saturation concentration $K_r$ for each reaction (chosen independently for forward and reverse reactions), and by modifying mass-action kinetics the following way:

- Monomolecular: $f_r=k_r\{[A_i]/(1+[A_i]/K_r)\}$

– Bimolecular: $f_r=k_r\{[A_i]/(1+[A_i]/K_r)\}\{[A_j]/(1+[A_j]/K_r)\}$

where $K_r=K_{avg}10^{random(-p/2,p/2)}$, with p the spread of saturation concentration counted in orders of magnitude. Saturating kinetics tend towards mass-action kinetics at low concentrations, as should be expected.

Chosen numerical values were $k_{avg\_monomolecular}=10^2\,s^{-1}$, $k_{avg\_bimolecular}=10^4\,M^{-1}.s^{-1}$ (typical $k_{cat}$ and $k_{cat}/K_m$ for enzymatic reactions), and $K_{avg}=10^{-2}\,M$. Standard Gibbs free energies of formation were assigned as $G_i/RT=random(0, 15)$ so that the maximum change in Gibbs free energy approaches that of ATP hydrolysis. Similar qualitative behavior were obtained irrespective of (s, p) values over several orders of magnitude of spreads. Results presented in the following were obtained with s=p=0.

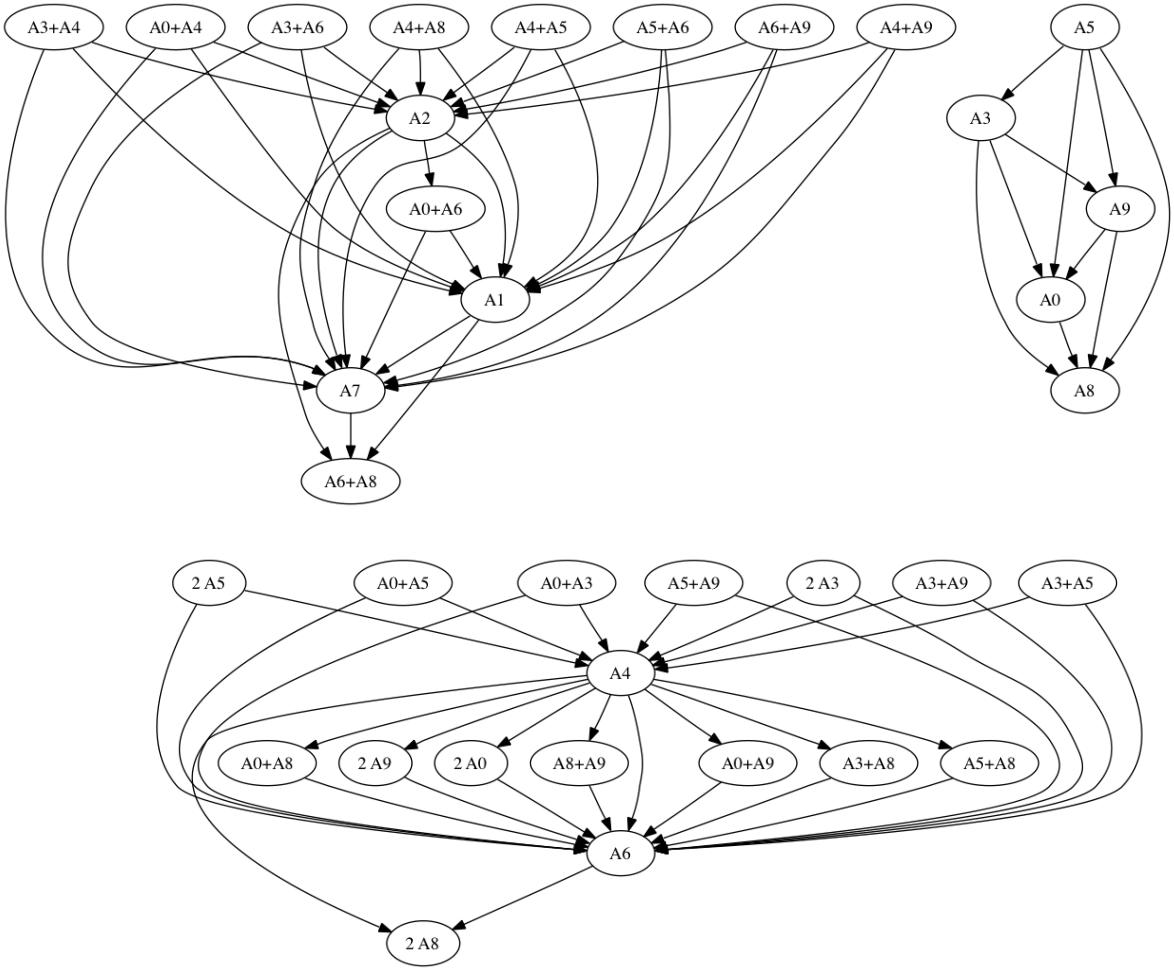



**Figure 1.** Example of a maximum-sized random conservative chemical reaction network, with N=10. Maximum size was reached for 74 direct reactions represented as arrows (148 reactions when counting both direct and reverse). Single mass vector components for this network are $\{m_i\}_{i=0,\ldots,9}=\{1, 3, 3, 1, 2, 1, 2, 3, 1, 1\}$.

# 3 Characterization of Random Reaction Networks

## 3.1 Equilibrium Behavior

The dynamics of the closed system are governed by the following equation:

$$dA/dt=Sf$$

where A is the N-vector of concentrations for the different chemicals, S is the NxR stoichiometry matrix, and f is the R-vector of reaction flux for the different reactions (components $f_r$ of vector f being a function of A through the kinetics given in the previous section).

Equilibrium concentrations were determined for different values of the system density D ($D=\sum_i m_i[A_i]$) by computing the concentration trajectories for sets of initial conditions with increasing concentrations. Unicity of the equilibrium for a given total density was checked by comparing computed equilibria under different initial conditions of same total density (equal distribution among chemicals vs. allocation to a single chemical, this process being repeated for each of the N chemicals).

Detailed balance at equilibrium is guaranteed from thermodynamic theory with mass-action kinetics [6], but not with saturating kinetics. Indeed, we observe that detailed balance is verified at any density with mass-action kinetics, but not for saturating kinetics: when the equilibrium deviates significantly from that with mass-action kinetics, some reactions are detailed-balanced while others are not.

Figure 2 gives equilibrium concentrations as a function of system density. At sufficiently low system density for which no reaction saturates, identical equilibrium concentrations are obtained with mass-action or saturating kinetics. At higher density, the behavior differs significantly with all the extra mass in the system going to a single chemical in the case of saturating kinetics.

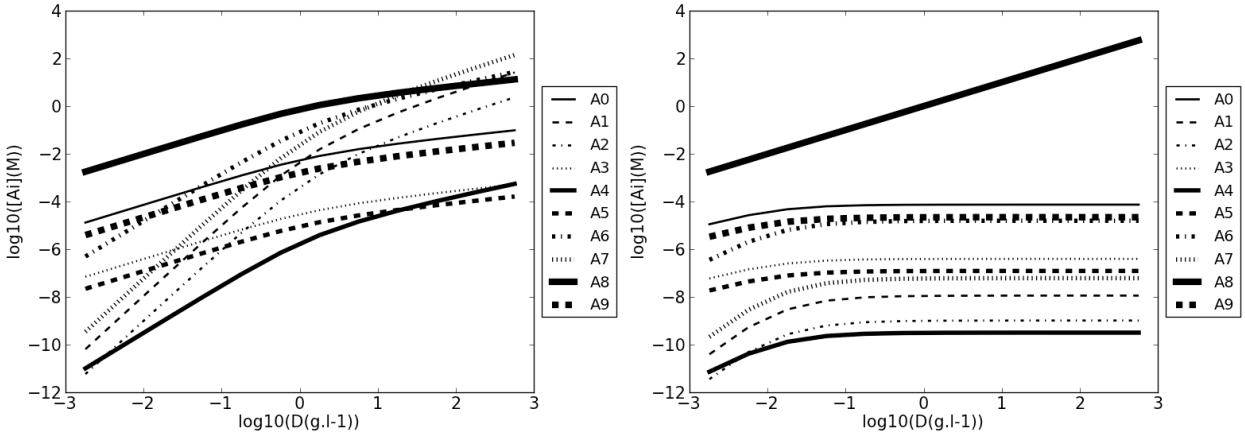



**Figure 2.** Equilibrium concentrations as a function of system density for an example randomly generated maximum-sized conservative system, with mass-action kinetics (left) and with saturating kinetics (right).

## 3.2 Behavior under External Nutrient Flux

We now consider system behavior when submitted to an external nutrient flux, $f_{nu}$, of any one of the N different species. The dynamics of the system are governed by the following equation:

$$dA/dt=Sf+f_{nu}$$

where $f_{nu}$ is the N-input flux vector having all components null except for the one injected chemical.

The rate of mass increase is conserved because the system is conservative: $m^TS=0 \Rightarrow m^TdA/dt=m^Tf_{nu}$, or equivalently: $\sum_{j=0,\ldots,N-1} m_j d[A_j]/dt=m_{nu}f_{nu}$.

We first consider saturating kinetics. Figure 3 shows a typical trajectory with $A_5$ arbitrarily chosen as nutrient. While dynamics on a short time scale (left) exhibit complex behavior, on a longer time scale (right) all chemicals except $A_8$ asymptotically reach constant non-zero concentrations, while all the injected mass is transformed into $A_8$, with $[A_8]$ diverging linearly asymptotically. This asymptotic behavior is independent of initial conditions. Using different nutrients leads to the same preferential transformation into $A_8$. In essence, the system acts as a directed transformation machine.

To further investigate this behavior, we have computed trajectories over a wide range of input flux, and up to a sufficiently large fixed given time. Figure 4 shows on log scales the values of densities $m_i[A_i]$ (top) and of their derivatives $m_i d[A_i]/dt$ (bottom) after the fixed given integration time.

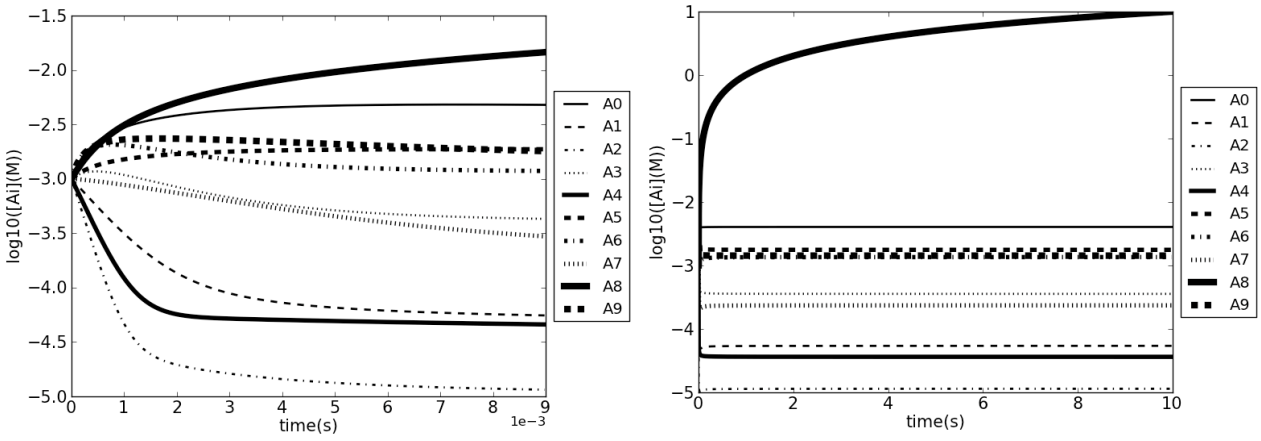



**Figure 3.** Concentrations vs. time on short (left) and long (right) time scales with a constant input flux $f_{nu}$=1 M.s$^{-1}$ of nutrient $A_5$ , for saturating kinetics. Initial conditions are 1 mM for all species.

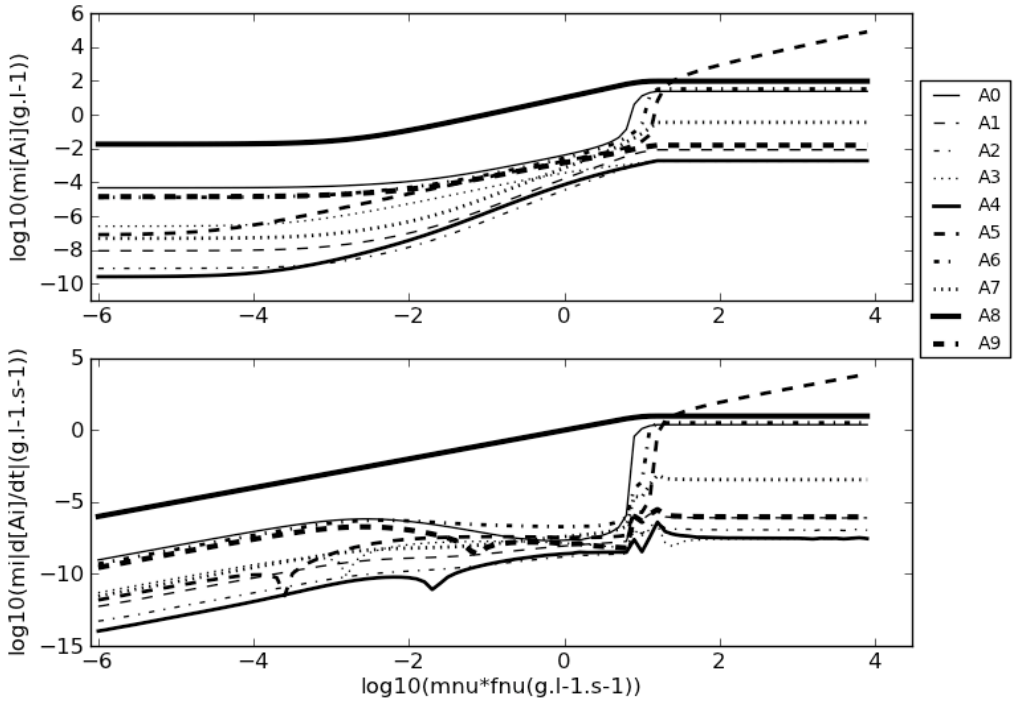



**Figure 4.** Densities $m_i[A_i]$ (top) and their derivatives $m_i d[A_i]/dt$ (bottom) after 10 s integration time, vs. nutrient density flux, for saturating kinetics. Nutrient is $A_5$. Initial conditions are 1 mM for all species.

Figure 4 (top) shows that at low $f_{nu}$ the injected mass is too low to significantly change equilibrium concentrations, even after the 10 s integration time. Above some threshold $f_{nu}$ (that depends on the system density, and thus on the integration time), $A_8$ becomes significantly more abundant. Consistently, Figure 4 (bottom) shows that at low $f_{nu}$ the conserved rate of mass increase $m_{nu}f_{nu}=\sum_{j=0,\ \ldots,\ N-1}m_j d[A_j]/dt$ is distributed among chemicals, while above the threshold $f_{nu}$ it nearly all goes to $A_8$. In this regime, asymptotical trajectories are typically as shown on Figure 3: $A_8$ diverges linearly and all other chemicals including the injected nutrient reach constant non-zero values.

The same directed transformation behavior as shown on Figures 3 and 4 is observed independently of the chosen nutrient, the one particular preferentially synthesized chemical ($A_8$ in the present example) being an intrinsic property of the system.

At even higher $f_{nu}$, a different regime is reached with the transformation capacity of the system being saturated, and all the extra added mass remaining intact. In this saturated regime, most reactions are saturated, and all reaction rates are locked independently of the $f_{nu}$ value.

When using mass-action instead of saturating kinetics, the same threshold behavior is observed, but all concentrations increase with increasing $f_{nu}$, and as expected no saturation regime is reached.

We have generated and characterized tens of random maximum-sized conservative systems, and observed that they all qualitatively behave the same way. The preferentially synthesized species depends on both network topology and kinetics, and is statistically a heavier lower-energy chemical.

## 4 Proto-Cell Model

We have found that most large conservative reaction networks with saturating kinetics effectively function as directed transformation machines, preferentially converting any nutrient into one particular chemical over a wide range of input fluxes. This particular chemical is an intrinsic property of the reaction network and is also the most abundant chemical in the system.

Most abundant chemicals in actual biological systems are typically structural molecules, first of which membrane molecules or their precursors. It is thus tempting to assign such a role to this most abundant and preferentially synthesized species in our model, which we will denote $A_{me}$.

### 4.1 Membrane Model

In the following, we assume that $A_{me}$ meets the membrane requirements: (i) ability to self-assemble in a continuous membrane and (ii) permeability of the self-assembled membrane to nutrients $A_{nu}$.

Regarding (i), we further assume $A_{me}$ is incorporated into the growing membrane at a molar rate per unit area: $\mathcal{I}_{me}=\mathcal{K}_{me}[A_{me}]/(1+[A_{me}]/K_{me})$ with $\mathcal{K}_{me}$ being a kinetic rate per unit area and $K_{me}$ a saturation concentration. This results from the unidirectional reaction $A_{me} \rightarrow A_{me\_structured}$ (once assembled, membrane constituents no longer react with other chemicals). Molar rate per unit volume $f_{me}$ (resp. kinetic rate $k_{me}$ per unit volume) are simply derived multiplying $\mathcal{I}_{me}$ (resp. $\mathcal{K}_{me}$) by the (*Area*/*Volume*) ratio: $f_{me}=k_{me}[A_{me}]/(1+[A_{me}]/K_{me})$.

As new $A_{me\_structured}$ get incorporated in the membrane, the membrane *Area* increases at a relative rate equal to $\mathcal{I}_{me}$ divided by the number of molecules per unit membrane area Nmea, that is an intrinsic property of the self-assembling molecules: $(1/Area)dArea/dt=\mathcal{I}_{me}/Nmea$.

By definition, cellular growth rate $\mu$ is the relative rate of volume increase, and is directly related to the relative rate of area increase by a multiplying factor (3/2 for a sphere, 1 for a filament). Assuming a filament shape for simplicity gives: $\mu=\mathcal{I}_{me}/Nmea=f_{me}/[membrane]$, with $[membrane]=Nmea(Area/Volume)$ being the effective concentration of structured membrane constituents if they were all dissolved in the cell volume.

Regarding (ii), we assume nutrients $A_{nu}$ can diffuse into the cell through a saturating process with rate: $f_{nu}=\mathcal{D}([A_{nu}]_{outside}-[A_{nu}])/(1+[A_{nu}]_{outside}/K_{nu\_outside})$, where $\mathcal{D}$ is an effective diffusion constant (in $s^{-1}$), $[A_{nu}]_{outside}$ is the nutrient concentration in the growth medium outside the cell, and $K_{nu\_outside}$ is a saturation concentration.

### 4.2 Sustained Cellular Growth

The dynamics of the system are now governed by the following equation:

$$dA/dt=Sf+f_{nu}-f_{me}-\mu A$$

where $\mu=f_{me}/[membrane]$ is the growth rate, and where the term $-\mu A$ represents the dilution factor (concentrations are reduced as cell volume grows). We are now interested in finding if such a system can sustain cellular growth, i.e. if there exists trajectories such that asymptotically dA/dt=0.

Our initial exploration of the ([membrane], $\mathcal{D}$, $K_{nu\_outside}$, $k_{me}$, $K_{me}$) membrane parameter space suggests the following: neglecting $f_{me}$ saturation (i.e. setting $K_{me}$ arbitrarily large), any numerically accessible parameter set leads to sustained steady-state growth once $[A_{nu}]_{outside}$ is sufficiently large, and nutrient flux is then independent of cytoplasmic state (i.e. $[A_{nu}]_{outside}>>[A_{nu}]$ ).

Furthermore, there is a broad parameter subspace for which the resulting cytoplasmic state $\{[A_i]\}$ is close to equilibrium. This corresponds to situations where $f_{nu}$<<metabolic rate (effective transformation rate of $A_{nu}$ in $A_{me}$ by the reaction network). As equilibrium concentrations are an intrinsic property of the system, the cytoplasmic state then only depends on the injected mass flux and is independent of which particular nutrient is injected provided its metabolization rate remains much faster than its injection rate. Figure 5 shows system behavior for such an example parameter set. Cytoplasmic concentrations were found to be close to equilibrium concentrations in the entire $[A_{nu}]_{outside}$ range.

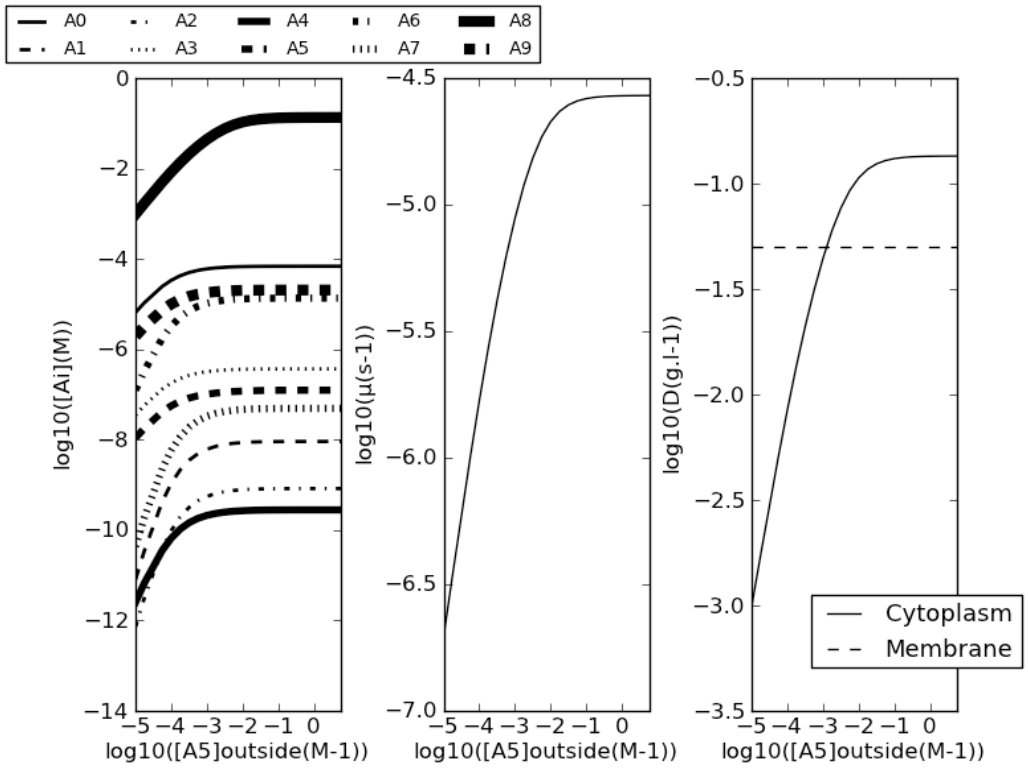



**Figure 5.** Steady-state cytoplasmic concentrations (left), growth rate (center), and cytoplasmic vs. membrane density (right), vs. outside nutrient concentration $[A_5]_{outside}$. Membrane parameters are $\mathcal{D}=10^{-3}\ s^{-1}$, $K_{nu\_outside}= 5\times10^{-3}$ M, $[membrane]=5\times10^{-2}$ M, $k_{me}=10^{-5}\ s^{-1}$. Steady-state concentrations are close to equilibrium.

## 5 Discussion and Conclusion

In the range of membrane parameters for which cytoplasmic concentrations are close to equilibrium, our proto-cell is essentially a machine that densifies matter, i.e. that converts a high-energy dilute growth medium (e.g. glucose) into a dense low-energy soup/paste enclosed within a growing membrane. Within a broad range of membrane characteristics, the cytoplasmic soup/paste may remain close to equilibrium, albeit at a larger density than the outside growth medium. Being close to equilibrium grants significant homeostasis: the cytoplasmic state only depends on the injected density flux, and is independent of the particular nutrient provided it is metabolized sufficiently fast.

Thus the growing cell seems to oppose the general principle of diffusion that would tend to equalize concentrations and densities. The apparent contradiction with thermodynamics arises from the fact that we have implicitly assumed that different reactions occur outside vs. inside the cell (glucose is very stable outside the cellular environment). This uncontested observation does not explain how this difference may have arisen in the first place, and this simply points towards the various origin-of-life scenarios [7].

In conclusion, we have shown that complex random chemical reaction networks effectively function as directed transformation systems, producing large amounts of a specific chemical (property of the chemical reaction network) when fed with any nutrient. Assuming this specific chemical can act as a precursor to a structured self-assembled membrane, we have built a proto-cell model capable of sustaining homeostatic cellular growth within a wide range of membrane parameters. In essence, large random conservative chemical reaction networks enclosed within their membrane behave as autopoietic systems, similar to random chemotons but without requiring any explicit informational template [7].

Many aspects of this work require further investigation. In particular, it is presumably the complexity conferred by the maximum size of our random conservative networks that leads to such deterministic and robust qualitative behavior. Characterization of systems below maximum size and as a function of a complexity metric such as R/N would be necessary to verify this point. Formal mathematical demonstrations of computationally observed behavior and/or the development of mean field approaches would also be necessary to extend our conclusions to very large N and R typical of living systems, and that are still beyond computational reach.

## Acknowledgements

We thank Pierre Legrain, Laurent Schwartz and Samuel Bottani for stimulating discussions and advice.

## References

[1] K. Kaneko, *Life: An Introduction to Complex Systems Biology*, Springer-Verlag, Berlin Heidelberg, 2006.

[2] A. Awazu and K. Kaneko, Ubiquitous “glassy” relaxation in catalytic reaction networks. *Phys. Rev. E*, 80:041931(7), 2009.

[3] Y. Kondo and K. Kaneko, Growth states of catalytic reaction networks exhibiting energy metabolism. *Phys. Rev. E*, 84:011927(8), 2011.

[4] H.X. Zhu, G. Rivas, and A.P. Minton, Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.*, 37:375-397, 2008.

[5] S. Schuster and T. Höfer, Determining all extreme semi-positive conservation relations in chemical reaction systems: a test criterion for conservativity. *J. Chem. Soc. Faraday Trans.*, 87:2561-2566, 1991.

[6] L. Onsager, Reciprocal relations in irreversible processes. I. *Phys. Rev.*, 37:405-426, 1931.

[7] P.L. Luisi, *The Emergence of Life: From Chemical Origins to Synthetic Biology*, University Press, Cambridge, 2006.

# Drought stress gene regulatory network reconstruction in Sunflower based on dynamical response to hormonal regulations

Gwenaëlle Marchand[1], Vân Anh Huynh-Thu[2], Sandrine Arribat[3], Didier Varès[1], David Rengel[1], Sandrine Balzergue[3], Patrick Vincourt[1], Pierre Geurts[2], Matthieu Vignes[4] and Nicolas B. Langlade[1]

[1] INRA, LIPM, Castanet-Tolosan, France
{gwenaelle.marchand,nicolas.langalde}@toulouse.inra.fr
[2] Université de Liège, Systems and Modelling, Liège, Belgium
[3] INRA, URGV, Evry, France
[4] INRA, MIA-T, Castanet Tolosan, France
matthieu.vignes@toulouse.inra.fr

**Keywords** Gene regulatory networks, abiotic stress, hormonal regulation, time-series gene expression, Sunflower, random forests, Gaussian graphical model, network fusion.

## 1 Background

Drought stress responses in plants are mediated through hormonal signals and leads to transcriptomic modifications [6,5]. If we see the plant as a complex system, hormones act in interaction, share transcriptomic targets and therefore signaling pathways [7]. In this framework, the use of gene regulatory networks (GRN) to represent such knowledge seemed a fruitful option: direct causal edges in the network represent, at the gene level, most interactions between key components, which occur at both cellular and molecular levels during the application of a stress to a plant [4]. To reconstruct a GRN of drought stress, we first adapted and validated (on simulated data) network reconstruction methods adapted to time-course gene expression data collected for different treatment (i.e. hormones). Then we applied them on real data collected for the 9 classes of plant hormones on Sunflower in time-course, on a selection of circa $150$ genes. Biological findings supported the biological relevance of many inferred relationships.

## 2 Methods

### 2.1 Data description

Experimental data were obtained from leaves of Sunflower (Helianthus annuus) plantlets, genotype (XRQ). During the development, nine different hormonal treatments and a control were applied to induce a response of the plant at time $t = 0$. Then, for each treatment, samples were collected and immediately frozen for 7 different times (0 i.e. just after treatment, 1, 3, 6, 9, 24 and 48h). For each time $\times$ hormone condition, 3 biological replicates were produced. Candidate genes were selected on 2 criteria of differential expression (assuming it stands for co- regulation): either as a response to ABA treatment (in particular in A. thaliana, the annotation of which being much richer than that of H. annuus) or in hydric stress conditions in controlled (greenhouse) or natural (field) environments. A list of $145$ genes was finally and analyzed by quantitative RT-PCR in each time $\times$ treatment condition, with standard checks and pre-processings. From there, we consider data organized in a $63 \times 145$ matrix $\mathbf{x}$ (see Section 2.3.1 below)

## 2.2 Simulated data

In order to assess the accuracy of the method we proposed to reconstruct a GRN from the Sunflower data at hand, we simulated simple yet plausible data sets. We only describe one of them here; their structure is very similar to that of the Sunfower data. First, a reference directed network was chosen. From this reference network, 9 different networks were derived by adding/removing/reversing 10, 20 or 30% of the edges for 3 networks in each case. Then, gene expression measures were simulated, the topology of which governed the dynamics of the system. In this framework, the true structures of the different networks are known and we could evaluate the accuracy (e.g. in terms of precision and recall) of the proposed inference strategy. Obviously, we do not claim that the generated data are even close to a biological data set because many assumptions (e.g. the artificial network structure, the Gaussian distribution of noises ...) are probably violated in real settings. However, we believe it was a necessary step to justify the potential for deciphering true causal relatioships from the Sunflower data.

## 2.3 Gene regulatory network formal description and inference methods

**2.3.1 Problem definition** We address the problem of recovering GRNs from time series expression data. The targeted GRNs are directed graphs with $p$ nodes, where each node represents a gene, and an edge directed from one gene $i$ to another gene $j$ indicates that gene $i$ (directly) regulates the expression of gene $j$. We only consider unsigned edges; when gene $i$ is connected to gene $j$, the former can be either an activator or a repressor of the latter.

In this paper, we assume that we have at our disposal $n = 9$ datasets $D_k (k = 1, \ldots, n)$. We assume that these datasets are respectively obtained from $n$ different perturbations of a system governed by an underlying GRN that specifies the plant response to an abiotic stress. Each perturbation corresponds to the induction of a specific hormone as described above. Each dataset $D_k$ contains gene expression levels measured at $T = 7$ different time points following a perturbation $k$:

$$D_k = \{\mathbf{x}_{k,1}, \mathbf{x}_{k,2}, \ldots, \mathbf{x}_{k,T}\}, \tag{1}$$

where $\mathbf{x}_{k,t} \in \mathbb{R}^p, t = 1, \ldots, T$ is a vector containing the expression values of all $p = 145$ genes at time point $t$ ($x^\top$ is the transpose of $x$):

$$\mathbf{x}_{k,t} = (x^1_{k,t}, x^2_{k,t}, \ldots, x^p_{k,t})^\top. \tag{2}$$

From these $n$ datasets, our goal is to learn $n + 1$ GRNs: one GRN resulting from each perturbation and a global consensus GRN taking into account all the perturbations. Two inference methods are considered in this paper, based on Random Forests (Section 2.3.2) and Graphical Gaussian Models (Section 2.3.3). They were combined to achieve GRN predictions.

**2.3.2 GRN inference with random forests** We extended a method called GENIE3 [11], which is based on random forests (RF, [3]) and that was originally proposed for the inference of GRNs from steady-state expression data. Random forests is here used as an ensemble of regression trees and the final prediction is an average combination of each tree prediction. Two random components are introduced in the construction of each tree: a "bagging" of the samples, which are divided in learning and test sets and a random selection of features to be used at each split of the tree. As in the original GENIE3 procedure, the problem of recovering a network of $p$ genes is decomposed into $p$ feature selection subproblems, where each of these subproblems consists in identifying the regulators of one gene of the network. In the presence of time series data, we make the assumption that the expression of each gene of the network at time point $t + 1$ is a function of the expression of the other genes of the network at the preceding time point $t$. Denoting by $\mathbf{x}^{-j}_{k,t}$ the vector containing the expression values at time point $t$ of all genes except gene $j$, we thus write:

$$x^j_{k,t+1} = f_j(\mathbf{x}^{-j}_{k,t}) + \epsilon_{k,t}, \forall k, t, \tag{3}$$

where $\epsilon_{k,t}$ is a random noise and functions $f_j$ only exploit the expression in $\mathbf{x}^{-j}$ of the genes that directly regulate gene $j$ in the underlying network. Recovering the regulatory links pointing to target gene $j$ thus

amounts to finding those genes whose expression at time $t$ is predictive of the expression of the target gene at time $t + 1$.

As in GENIE3, our procedure exploits feature importance scores derived from RF models to rank candidate regulators of each gene.

1. a RF model is trained to predict the expression of the target gene at time $t+1$ (*i.e.* $x_{t+1}^{j}$) from the expression levels of all other genes at time $t$ (i.e. $\mathbf{x}_t^{-j}$).
2. candidate regulators are ranked according to variable importance scores derived from the RF model. Importance scores are computed as the total variance reduction due to splits based on the corresponding regulator expression, averaged over all nodes and trees in the forest [1]).
3. a global ranking of all regulator-target gene edges is obtained by merging all individual target gene rankings with their associated importance scores and a network prediction is obtained by thresholding these scores.

RF importance scores are not statistically interpretable, which makes difficult the determination of an importance threshold to obtain a single and interpretable network prediction. We therefore propose to replace these scores by a new score that can be interpreted statistically.

To compute this score, we add to the dataset an artificial $(p + 1) = 146$th random gene, whose expression values are obtained by randomly permuting the $n \times T$ expression values of a gene randomly selected among the $p = 145$ original genes (making the new gene uncorrelated to all other genes). We then run the RF learning procedure described above to obtain a ranking of the GRN edges, including edges involving the random gene. We repeat this process $1,000$ times and take as the score of a GRN edge, the proportion of the $1,000$ rankings where this edge was ranked above all the edges involving the random gene. The resulting edge score is then interpreted as the probability that this edge is ranked by the RF modelling at a higher level than a spurious edge (the higher, the better).

The previous procedure can be applied separately on each time series $D_k, k = 1, \ldots, n$ or on the union of all time series to obtain respectively the $n$ perturbation-specific GRNs and the global consensus GRN. However, each individual time series being rather small and expecting only limited differences between these networks, we preferred the following procedure to obtain the $n$ perturbation-specific GRNs: first, RF models for all genes are trained on the union of all time series. Then, pertubation-specific importance scores are obtained by re-propagating the instances from each dataset $D_k$ separately into these RF models and re-computing variable importance scores only from these instances.

**2.3.3 Inferring multiple GRN structures with Gaussian graphical models** We used here a Gaussian Graphical Modelling (GGM), a widely used statistical tool for the reconstruction of networks of regulatory relationships between genes. The main difficulty stands in the high-dimensionality of the data: the number of variables (genes) exceeds the number of samples (combination of treatment $\times$ time point). If samples are considered as independent, each microarray is considered as the observation of multivariate Gaussian random variables. Its intrinsic dependencies are encoded in the associated covariance matrix or more precisely in the inverse of this matrix: the precision matrix. In fact, non-zero entries of the precision matrix fully determine non-independent couples of variables in the network. Because of the high-dimensionality of the data set, we chose to rely on the widely used lasso [2], an $\ell_1$ penalization technique, which basically assumes sparsity of the network topology. More precisely, using notations previously introduced, we first assumed a first-order auto-regressive model on centred data in each condition $k$:

$$x_{k,t} = x_{k,t-1} A_k + \epsilon_{k,t},$$

where matrix $A_k$ contains the effects of all genes at time $t - 1$ onto genes at time $t$. This modelling is close to that of random forests of Equation 3.

If we treat the case of one hormonal treatment and omit subscript $k$, maximizing the log-likelihood of the model is equivalent to the following optimization problem: $\max_A \left\{ \mathrm{Tr}(^\top V\, A) - 1/2 \mathrm{Tr}(^\top A\, S\, A) \right\}$ and the

solution (maximum likelihood estimator) is given by $\hat{A}^{MLE} = S^{-1}V$, where we denoted by $S$ the empirical variance-covariance matrix and by $V$ the empirical temporal covariance matrix (see [9]) and we omitted subscript $k$ which referred to the hormonal treatment. An $l_1$ penalty on matrix $A$ which encodes non-zero coefficient of the auto-regressive model was used to circumvent the high-dimensionality of the problem ($S$ being not invertible) under sparsity assumptions on non-zeros elements of matrix $A$.

In our framework, $9$ different conditions, which correspond to $9$ different matrices $(A_k)_{k=1...9}$, have to be considered. If we ignored the relationships between the different hormonal treatments, we would simply optimize a problem which would be the sum of $9$ problems similar to the one of the previous paragraph. We instead combined the temporal approach of [9] to the multiple graph structure inference scheme of [12], which is written for iid Gaussian graphical models. In our context indeed, samples are not independent: time-course experiments imply that temporal patterns of expression govern the observed regulatory relationships between genes. We used a so-called "intertwined" estimation of matrices $A_k$'s. It renders the model parameter estimation over different hormonal conditions not separable anymore. More precisely, the objective function (the log-likelihood) is slightly modified and instead of using the $9$ matrices $V_k$ and $S_k$ separately, we used a convex combination that account for a part which is specific to the hormonal treatment and the other part which is a mean of each matrix over all conditions. The mixing parameter of the convex combination is arbitrarily set to $1/2$; if it were equal to $1$, all data sets would be pooled as a single one and if it were set to $0$, the estimate of the matrices corresponding to each hormonal treatment would become independent. We restricted the number of edges in each network to be no more than $200$ for computational reasons.

The method, implemented in the R package `SIMONe` [10], proposed an integrated structural organization of the network through an on-line latent clustering of genes. We did not exploit this refinement, but to obtain an initial network through a burn-in like run before running the algorithm. Lastly, we made the prediction over edges from matrices $A_k$ more robust by applying the method we just described $200$ times on bootstrapped version of the samples, in the spirit of the bootstrap lasso introduced in [8]. A time-series length was first uniformly chosen between 3 and 7 and then time points were picked up at random for each treatment,with the same time series length preserving the time ordering, so that different response times to different hormones could be considered. An edge was identified as be significant when it was predicted in $80\%$ of the bootstrapped runs of the algorithm. An edge is supported by $80\%$ of the bootstrapped replicates but not necessarily the same for all edges. The rationale behind this heuristics is that we prefered to focus on edges that appear in most bootstrapped repeats of the algorithms but in possibly varied contexts for each edge.

Lastly, note that this approach does not directly produce a consensus network. We chose to give a weight to each edge of the created global network equal to the median of the same edge for all hormone networks. Hence if an edge is absent of most networks, it won't appear in the global GGM network. On the contrary, if it receives a large weight in most hormone networks, it will by one of the major edges of the global GGM network.

**2.3.4 Fusion of RF and GGM networks** At this point, for each treatment and for a global consensus network, we have two lists of edge weights. At the end of the day, we want to end up with one such list of "optimal" length for each treatment.

We dealt with the optimality point by setting a threshold of validation for an edge in over $20\%$ of the bootstraps for the GGM approach and having a threshold of $40\%$ of edge significance in regards of a random gene for the RF approach.

We used the results of the two methods to build a robust consensus network. For each hormonal network and for the global network we considered that an edge was robust enough to be represented in the consensus network if it was simply selected by both GGM and RF methods. The final consensus network is formed by the union of all these robust edges. For example [14] observed that integrating several approaches seems beneficial in a different settings. citedream-wisdom even established the consistency of the so-called community prediction paradigm, when the number of (independent) inference methods becomes very large, given they only achieve performance better than random guessing.

# 3 Results

## 3.1 Simulated data

We here quickly compare the 2 methods presented in Sections 2.3.2 and 2.3.3. Figure Fig. 1 (left-hand side) represents the precision versus recall curve obtained by the two inference methods we proposed in the case of one of the simulated networks (of size 100 genes and comprising 200 edges). The precision is the proportion of correctly predicted edges among all predicted edges; the recall is the proportion of all correctly predicted edges among all edges to be predicted (hence 200 here). The curve is obtained by setting different thereshold values for edges significances. Note that edges significance in one method cannot be compared to the one in the other approach and they have to be set independently.

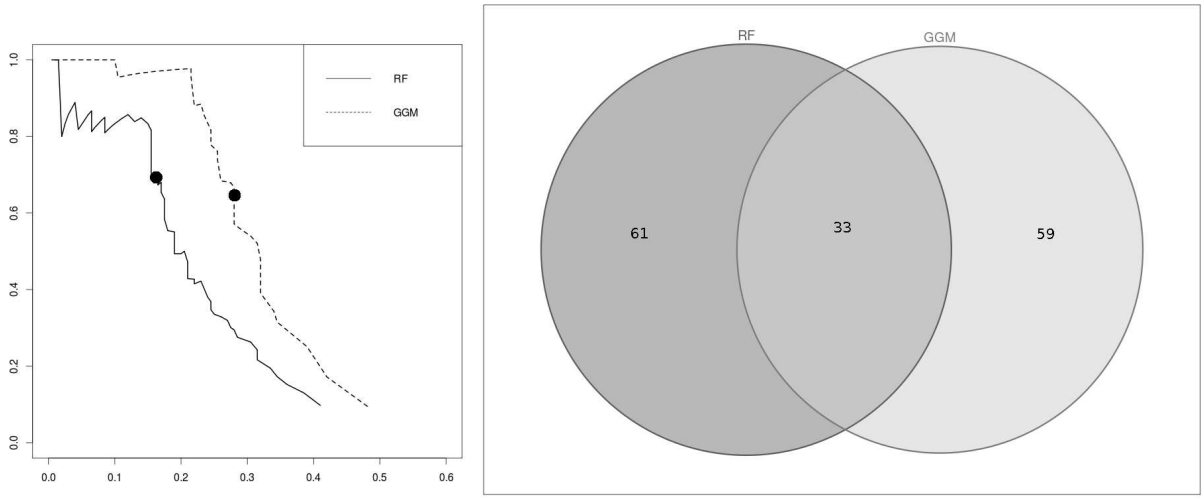



**Figure 1.** (Left) Precision versus recall curve for the inference of a simulated network and (Right) Venn Diagram of the 2 predicted networks depicted by the 2 large dots on the left (see text)

Both approaches perform reasonnably well. For example, if we set confidence threshold to $0.4$ and $0.2$ for respectively the RF and GGM appraoches, we obtain networks of size $94$ and $92$ edges respectively. RF achieves a precision of $0.673$ and a recall of $0.165$, whislt GGM has a slightly lower precision of $0.615$ but an interesting recall of $0.28$. It must be noted that none of these two appraoches is able to identify more than about $40\%$ of the network edges (i.e. $80$ over $200$) with a reasonnable level of precision ! The fact that the GGM approach performs better is slightly artificial and is probably due to the simulated Gaussian noise for the data set. Moreover, the RF approach performs quite comaprably to the GGM on the global network reconstruction (the retained Gold standard network is then the union of all hormone networks).

The networks identified by the two large dots on the precision vs recall curve are then compared on the right-hand side of Figure Fig. 1. This Venn diagram shows that although both methods have similar performances, they do not focus on the same predictions. They only share 33 edges: it is less than half of their predictions. The very interesting point here is that all these 33 edges are correct edges. Hence combining the 2 approaches allows us to get a $100\%$ precision for a recall of $16.5\%$, a score that was not achieve by any of the 2 methods. For example, setting the threshold to $0.25$ and $0.1$ gives an intersection of $38$ edges and $37$ among them are correct (precision is then $97.4\%$ and recall $18.5\%$). Given the improvement is relatively weak, we prefer to keep a conservative approach where we achieve acceptable performances for both methods individually and their fusion is very good.

We are aware that such precision cannot be achieved on real data sets because of quite different patterns observed in real data sets and which we have not accounted for. We then expect performances (given they could be defined) to deteriorate. However, we have a high confidence in the first edges which are confirmed by both methods. In the next Section, we present the results we obtained on the Sunflower data set and which confirmed this assertion.

### 3.2 A Sunflower network in response to hormonal treatment with discussions

After combining the results of the two methods of inference, we obtain on Sunflower data a network with $69$ nodes connected by $79$ edges. The topology of this network is characterized by few genes which are very highly connected and a majority of genes with a low connectivity. This scale-free topology is characteristic of biological networks. The genes with a low connectivity have Gene Ontology (GO) terms related to metabolism, on the contrary to genes with a high connectivity that have GO terms related either to DNA binding and transcription factor or to anionic transport through the cell membrane.

In addition to this topology characteristic, we validate our Sunflower network using external data from the plant model Arabidopsis thaliana. These data are composed of seven hormonal treatments with (only) 3 time points. Due to this difference in the dynamics sampling, we were not able to define a network from them, which would play the role of a Gold(ish) Standard. We looked for gene expression correlations that confirm conclusions drawn from Sunflower data. Among the $116$ Arabidopsis genes homologous to the $145$ Sunflower genes in our experiments, significant correlations between connected gene pairs are much more frequent than between unconnected gene pairs, according to an hypergeometric test ($\mathrm{p-value} = 0.005$).

The inferred Sunflower network inferred has two hubs encoding putative nitrate transporters. Arabidopsis mutants for these genes show stomatal closure defects due to the modification of the osmotic potential in guard cells. This result suggests that the nitrate transport in guard cell of stomata plays an important role in drought stress regulation and responses in Sunflower. In the Sunflower network, these two genes share several target genes but no source genes, and represent two different pathways for the same drought stress responses.

Finally GRNs constrain genetic variability and evolution of composing genes. According to evolutionary theories, highly connected genes are subjected to important trade-offs limiting their natural variability. Therefore, contrasting evolution forces are reflected in the network topology. Looking for evidences of these theories, we searched for correlations between topological parameters of the network and genetic diversity statistics within and between five populations of Sunflower relatives (H. petiolaris, H. argophyllus, wild H. annuus, landraces and elite lines of H. annuus). Significant correlations between genetic diversity statistics (e.g. Fst) and network topology descriptors (e.g. eccentricity or average shortest path length) were observed for several populations. This allowed us to develop new hypothesis about the role of this gene regulatory network, controlling drought stress response, in the evolution of species such as H. argophyllus or during the domestication of H. annuus.

## 4 Conclusion

In this work, we introduced two GRN network inference methods, which were motivated by a specific biological data set. This data set consists of expression data of genes, which are believed to play a role in the response of the Sunflower plant to an abiotic stress, in relation to hormonal regulation. The dynamical expression measures were obtain under $9$ different hormonal treatments. We explored the potential of the two approaches, based on random forests or Gaussian graphical models on simulated data and in particular showed that combining the two approaches lead to much more valuable predictions. Lastly, we applied our reconstruction strategy to the initial Sunflower data and reported some of the biological meaningful insights, which we found.

Obviously, this work is not the end of the line; many developments could be foreseen. As advocated by [15], other types of inference methods could be considered in the network fusion process. Additionaly, the robustness of our predictions could be checked in the case where genes which actually play a role in the system are not observed; it is anticipated that the importance of the missing genes (e.g. hub versus end product) and the number of missing genes will be critical aspects to take into account. Lastly, the generalization of such consensus approaches to other biological data sets might be an option to consider in order to achieve more robust and high quality predictions. For example, one might be tempted to go beyond the GRN reconstruction and to infer direct causal relationships between biological entities (genes, proteins, metabolites) which actually interact, in the same vein of [13] but with an automatic instead of manual network reconstruction. This would

be a nice research area marriage, where biologists and quantitative researchers will need to discuss the need of theoretical developments to answer crucial biological questions in a global system modelling framework.

## Acknowledgements

The authors are very grateful to C. Charbonnier, J. Chiquet, S. Lèbre, B. Mangin, M.-L. Martin-Magniette and A. Rau for fruitful discussions on this work.

## References

[1] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*, Chapman & Hall/CRC, Boca Raton, 1984.

[2] R. Tibshirani, Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58:267-288, 1996.

[3] L. Breiman, Random forests. *Machine Learning*, 45:5-32, 2001.

[4] E. Davidson and M. Levin, Gene regulatory networks for development. *PNAS*, 102:4935, 2005.

[5] J. Zhang, W. Jia, J. Yang and AM. Ismail, Role of ABA in integrating plant responses to drought and salt stresses. *Field Crops Research*, 97:111-119, 2006.

[6] K. Yamaguchi-Shinozaki and K. Shinozaki, Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annu Rev Plant Biol.*, 57:781-803, 2006.

[7] JL. Nemhauser, F. Hong and J. Chory, Different plant hormones regulate similar processes through largely nonoverlapping transcriptional responses. *Cell*, 126:467-75, 2006.

[8] F. Bach, Bolasso: model consistent Lasso estimation through the bootstrap. in WW. Cohen, A. McCallum, ST. Roweis (Eds.), *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML)*, ACM International Conference Proceeding Series 307, Helsinki, Finland, pp. 33-40, 2008.

[9] C. Charbonnier, J. Chiquet and C. Ambroise, Weighted-lasso for structured network inference from time course data. *Statistical Applications in Genetics and Molecular Biology*, 9, 2009.

[10] J. Chiquet,A. Smith,G. Grasseau, C. Matias and C. Ambroise, SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics*, 25:417-418, 2009.

[11] VA. Huynh-Thu, A. Irrthum, L. Wehenkel and P. Geurts, Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5:e12776, 2010.

[12] J. Chiquet, Y. Grandvalet and C. Ambroise, Inferring multiple graphical structures. *Statistics and Computing*, 21:537-553, 2011.

[13] C. La Rota, J. Chopard,P. Das,S. Paindavoine, F. Rozier,E. Farcot, C. Godin,J. Traas and F. Monéger, A data-driven integrative model of sepal primordium polarity in Arabidopsis. *Plant Cell*, 23:4318-4333, 2011.

[14] M. Vignes, J. Vandel, D. Allouche, N. Ramadan-Alban, C. Cierco-Ayrolles, T. Schiex, B. Mangin and S. de Givry, Gene regulatory network reconstruction using Bayesian networks, the Dantzig selector, the Lasso and their meta-analysis. *PLoS ONE*, 6:e29165, 2011.

[15] D. Marbach, JC. Costello, R. Küffner, NM. Vega, RJ. Prill, DM. Camacho, KR. Allison, The DREAM5 Consortium, M. Kellis, JJ. Collins and G. Stolovitzky, Wisdom of crowds for robust gene network inference. *Nature Methods*, 9:796-804, 2012.

# Bootstraping PPI networks to improve biological quality of partitions

Benoît ROBISSON[1], Alain GUÉNOCHE[2] and Christine BRUN[1]

[1] TAGC, UMR1090, Aix-Marseille Université, Campus de Luminy, Case 928, 13288 Marseille cedex 9 France
{robisson, brun}@tagc.univ-mrs.fr

[2] IML, CNRS, Aix-Marseille Université, Campus de Luminy, Case 907, 13288 Marseille Cedex 9, France
guenoche@iml.univ-mrs.fr

**Abstract** *The inherent uncertainty in biological data is a recurrent problem when analyzing biological networks such as protein protein interaction (PPI) networks. Here, inspired by classical methods in phylogeny, the bootstrap clustering adds uncertainty to the graph. This method generate altered networks similar to the initial one. A potential partition is then computed for each altered network and a consensus of partitions is made from this profile of potential partitions. We have evaluated this approach by generating a reference partition and its corresponding network. According to the Rand index corrected by chance, the consensus partition of altered networks (ANs) is closer to the reference partition than that of unaltered networks (Uns). We have also applied our method to PPI networks and have developed a functional homogeneity score to assess the biological relevance of the clusters obtained before and after the introduction of uncertainty. Strikingly, partitions of the ANs again scored better than partitions of the Uns. Our analysis shows that for sparse networks such as PPI, reasonable levels of alteration improve clustering results and lead to more meaningful biological clusters. These findings call for a reconsideration of the effect of uncertainty in protein interaction networks: what has long been considered a drawback could improve the results of clustering algorithms.*

**Keywords** network partitioning, protein-protein interaction network, noise

## Ajout d'incertitude dans les réseaux PPI et amélioration de la qualité biologique des partitions

**Résumé** *L'incertitude intrinsèque aux données biologiques est un problème récurrent de l'analyse des réseaux d'interactions protéine-protéine (PPI). Inspiré par une méthode classique en phylogénie moléculaire, le bootstrap clustering tend à y remédier. Cette méthode génère des réseaux altérés similaires au réseau initial, construit une partition potentielle pour chacun des réseaux altérés et, pour finir, calcule une partition consensus à partir de ce profil de partitions potentielles. Pour évaluer cette approche nous avons généré une partition de référence à retrouver à partir d'un réseau initial, et plusieurs réseaux similaires à ce réseau initial. D'après l'indice de Rand corrigé, le consensus des partitions obtenus par partitionnement des réseaux similaires est plus proche de la partition de référence que la partition issue du réseau initial. Nous avons appliqué cette méthode à des réseaux PPI et avons défini un score d'homogénéité fonctionnelle pour évaluer la pertinence des modules obtenus avant et après la procédure de bootstrap. De nouveau, les partitions consensus sont meilleures que les partitions initiales. Notre analyse montre que, pour les réseaux peu denses tels que les réseaux PPI, des niveaux d'altération raisonnables améliorent les résultats du partitionnement et produisent de meilleurs modules biologiques. Ces résultats amènent à une reconsidérer l'effet du bruit dans les réseaux PPI; ce qui a longtemps été considéré comme une faiblesse peut améliorer les résultats des algorithmes de partitionnement.*

**Mots-clés** partitionnement de graphe, réseau d'interactions protéine-protéine, bruit

## 1 Introduction

Les premiers réseaux d'interactions protéine-protéine (PPI) ou 'interactomes' sont apparus en 2000 [1], construits grâce aux résultats des cribles double-hybrides réalisés chez la levure (Y2H) [2]. Cette technique,

développée en 1989 pour détecter des interactions directes entre deux protéines, est maintenant classiquement utilisée pour les cribles à grande échelle [3]. Sur la base des résultats obtenus lors de ces cribles, la taille de l'interactome humain a été estimée à environ 130 000 interactions [4], alors que le plus grand réseau dont nous disposons actuellement ne contient qu'environ 80 000 interactions. Les réseaux d'interactions protéine-protéine (PPI) sont donc incomplets.

En effet, bien que théoriquement les interactomes représentent l'ensemble des interactions possibles entre toutes les protéines d'un organisme, la qualité et la couverture des interactions détectées par les cribles double-hybrides ont longtemps été sujettes à caution. Alors que certaines interactions existantes ne sont pas détectées (les 'faux-négatifs') d'autres, qui n'existent pas, le sont (les 'faux positifs'). Dans les réseaux issus de ces cribles, les interactions manquantes sont dues essentiellement à des raisons méthodologiques : soit (i) les conditions expérimentales du crible ne permettent pas de détecter certaines interactions, comme par exemple lorsque des modifications post-traductionnelles des protéines qui n'ont pas lieu dans la levure sont nécessaires à l'interaction, soit (ii) l'espace des interactions n'a pas été assez couvert car le crible n'a pas été reproduit un nombre de fois suffisant , soit (iii) toute les interactions possibles ne sont pas testées expérimentalement. Pour remédier au problème de couverture, il a été montré qu'il est nécessaire de répliquer un crible au moins 6 fois pour atteindre la saturation, c'est à dire pour détecter au moins 90 % des interactions présentes [4]. D'autre part, des interactions de type 'faux positifs' peuvent avoir des origines méthodologiques ou biologiques. Les 'faux positifs' méthodologiques sont dus aux propriétés auto-activatrices des protéines testées, qui peuvent activer le gène rapporteur sans nécessiter d'interactions avec un partenaire. Les 'faux positifs' biologiques correspondent soit à des protéines ayant des interactions peu ou pas spécifiques, soit à des protéines qui ne sont jamais en présence *in vivo*. Toutefois, il est important de noter que ces interactions, qualifiées de 'non physiologiques', n'en restent pas moins des interactions possibles biophysiquement et biochimiquement. Elles pourraient donc représenter l'expression du 'bruit' biologique intrinsèque au vivant, et leur appellation de 'faux positifs' pourrait n'exprimer que les limites de l'étendue de nos connaissances. Plus particulièrement, l'organisation du réseau d'interactions protéine-protéine pourrait refléter le bruit biologique comme une condition à la robustesse des systèmes biologiques.

Afin d'extraire de l'information biologique des grands réseaux PPI et d'identifier des sous-réseaux pertinents biologiquement, un grand nombre de méthode de partitionnement de graphe et d'analyse de réseaux ont été proposées (pour revue [5]). Cependant, plusieurs incertitudes et difficultés liées à l'état des connaissances sont rencontrées lors de l'évaluation et de la validation mathématique et biologique de ces méthodes : (i) les réseaux biologiques étant incomplets et bruités, le modèle de graphe sous-jacent utilisable par exemple pour des simulations n'est pas connu ; (ii) on ne connait pas de partition de référence pour un réseau biologique qui pourrait permettre l'évaluation de la pertinence des méthodes et algorithmes ; (iii) les données que l'on utilise à ces fins (complexes protéiques, groupes de protéines annotées au même terme Gene Ontology...) ne constituent pas une partition du graphe à proprement dit (certaines protéines appartiennent à plusieurs groupes, d'autres n'appartiennent à aucun d'entre eux).

Dans ce travail, nous nous sommes donc posés la question de la résistance des méthodes de partitionnement au bruit dans les réseaux. Quelle est l'influence du bruit dans les données biologiques sur la qualité des sous-réseaux identifiés ? Comment se prémunir d'effets possiblement néfastes à la qualité du partitionnement ?

Pour cela, les réseaux potentiels, proches du réseau étudié, sont explorés grâce à une méthode de bootstrap clustering [6]. Elle consiste à modifier raisonnablement, à plusieurs reprises, le réseau initial pour calculer un profil de partitions potentielles. L'ajout d'incertitude dans le réseau introduit une variabilité parmi les partitions potentielles, ce qui permet d'identifier les paires de protéines les plus robustes (les plus souvent réunies) et ainsi de calculer une partition consensus. La question est alors de savoir si cette partition consensus est plus pertinente biologiquement qu'une partition issue du réseau initial non modifié. Pour valider l'approche, des simulations ont été réalisées pour évaluer les performances de la méthode, dans lesquelles la partition attendue est connue. Pour répondre à la question, cette méthode de bootstrap clustering a été appliquée à des réseaux biologiques, et la pertinence biologique des partitions obtenues a été évaluée. Ces partitions calculées ont été comparées à des références biologiques, puis un score a été défini pour évaluer l'homogénéité fonctionnelle

des classes. Ces résultats montrent que le fait d'introduire de l'incertitude dans un réseau permet d'améliorer de façon significative le partitionnement de celui-ci.

## 2 Validation théorique

### 2.1 Bootstrap clustering

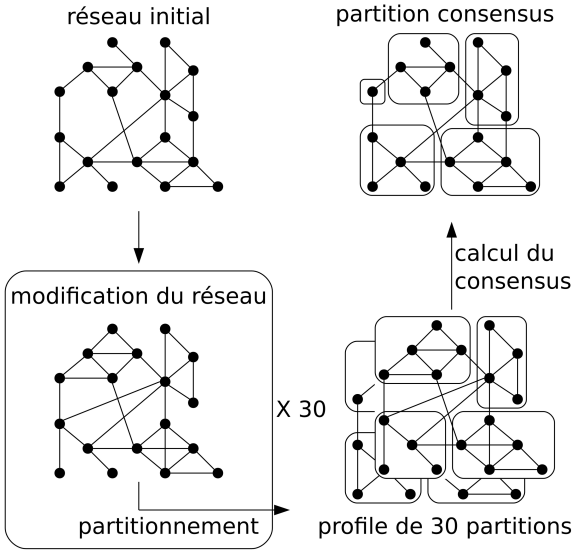



**Figure 1.** Étapes de la méthode de bootstrap clustering.

La méthode est similaire au bootstrap utilisé en phylogénie moléculaire pour mesurer la robustesse des arbres. La différence est qu'ici, le bootstrap clustering a pour but de construire une classification plus robuste. La méthode suit les étapes (Fig. 1) suivantes :

1. La génération de réseaux modifiés proches du réseau initial. Deux façons de modifier le réseau, donc d'introduire du bruit, ou de l'incertitude, sont étudiées :
   – type élongation : en modifiant la pondération, initialement fixé à 1, de toutes les arêtes avec un taux d'élongation *e*, ajouté ou soustrait aléatoirement aux arêtes,
   – type ajout-pondération en utilisant l'indice de Dice [7] ($D$) : on ajoute des arêtes aléatoires parmi les paires d'éléments $u, v$ qui ont au moins un sommet adjacent commun ($D(u, v) < 1$) avec une probabilité $a$ donnée, appelée taux d'ajout, et on pondère toutes les arêtes $(x, y)$ par $1 - D(x, y)$.
2. Le partitionnement des réseaux modifiés par l'algorithme TFit [6], pour Transfert-Fusion itéré, une méthode multi-niveaux qui optimise la modularité de Newman [8]. On réalise ainsi un profil de partitions potentielles.
3. Un algorithme de consensus de partitions calcule la partition médiane (consensus) des partitions potentielles.

Aucun des types de modification de réseaux ne retire d'arêtes car, dans les interactomes que nous étudions, les arêtes présentes sont relativement sûres. Le taux de faux positifs est très faible en regard du taux de faux négatifs. De plus, nos réseaux étant déjà très peu denses, éliminer des arêtes risque de déconnecter le réseau. Pour prendre en compte ce faible taux de faux positifs, nous diminuons la pondération de certaines arêtes au lieu de les retirer, selon le type de modification choisi.

Un protocole de simulations sur des petits réseaux aléatoires (200 sommets), générés à partir d'une partition source, a été réalisé par les auteurs. Dans ce protocole, le réseau construit à partir de la partition source est généré suivant un modèle d'Erdös-Rényi à 2 paramètres $d_i$, la probabilité de tirer une arête entre deux éléments appartenant à une même classe, et $d_e$, la probabilité de tirer une arête entre deux éléments appartenant à deux classes différentes. Pour les détails du protocole et les résultats, nous renvoyons le lecteur à la publication [6]. Ils aboutissent à la conclusion que la partition consensus est plus proche de la partition source que la partition initiale, calculée par TFit sur le réseau initial. C'est ce que nous souhaitons vérifier sur des réseaux aléatoires de taille et de densité comparables aux réseaux biologiques.

## 2.2 Simulations

Pour évaluer les performances du bootstrap clustering dans la recherche d'une partition robuste et pertinente, l'amélioration apportée par la partition consensus par rapport à la partition initiale est mesurée. Pour cela, le protocole de simulations a été étendu : la partition source est fixéee et les réseaux simulés de différentes tailles et de différentes densités sont générés et partitionnés, pour mesurer la proximité des partitions calculées (initiale et consensus) avec la partition attendue (source).

On travaille sur 2 types de réseaux :
– taille moyenne, $v = 2000$ éléments séparés en $n = 50$ classes équilibrées,
– grande taille, $v = 6000$ éléments séparés en $n = 100$ classes équilibrées.

Le nombre de classes est choisi de façon à ce que la taille des classes obtenues soit similaire à ce qui est attendu dans les réseaux biologiques. La densité, rapport entre le nombre d'arêtes présentes divisé par le nombre d'arêtes possibles, est un paramètre essentiel. Pour chaque type, on teste différentes densités (densité intra-classe $d_i$, densité inter-classe $d_e$), en cherchant à se rapprocher des réseaux biologiques qui sont très peu denses, tout en gardant la connexité du réseau (Table 1).

| | | 2000 / 50 | | 6000 / 100 | |
|---|---|---|---|---|---|
| $d_i$ | $d_e$ | densité | composante connexe | densité | composante connexe |
| .50 | .05 | .0589 | 1 | .0544 | 1 |
| .30 | .03 | .0353 | 1 | .0326 | 1 |
| .10 | .001 | .0029 | 1.04 | .002 | 1 |

**Table 1.** Description des réseaux simulés étudiés.

Pour tous les réseaux générés avec chacune de ces densités, TFit et le bootstrap clustering sont appliqués. Différents paramètres sont testés, tout d'abord avec la modification de type élongation avec plusieurs taux d'élongation *e* (0.1, 0.2, 0.3), puis avec la modification de type pondération-ajout avec plusieurs taux d'ajout *a* (0.3, 0.5, 0.7). Pour chacun de ces tests, l'indice de Rand corrigé [1] est mesuré entre la partition de référence et la partition initiale d'une part, et entre la partition de référence et la partition consensus d'autre part (Fig. 2 page ci-contre).

Ces résultats montrent que la partition consensus est presque toujours plus proche de la partition source (que l'on cherche à retrouver) que la partition initiale. Plus précisément, l'utilisation des modifications de réseau de type élongation produit des résultats toujours meilleurs et stables, par contre, avec les modifications de type pondération-ajout, les résultats sont plus instables, et généralement moins bons. Ces résultats encourageants montrent que le bootstrap clustering permet de retrouver une partition plus proche de la partition source que l'algorithme TFit sans bootstrap sur des réseaux simulés. Cependant, les performances de ces deux méthodes sont fortement dépendantes du niveau de difficultés des simulations.

## 2.3 Évolution des résultats selon la difficulté du problème

Dans les simulations utilisées, il est possible de choisir la difficulté du problème à résoudre par les algorithmes de partitionnement en sélectionnant les densités intra-classes et inter-classe. Lorsque la densité intra-classe est largement supérieure à la densité inter-classe, le problème est facile, mais lorsque ces deux densités sont proches, les algorithmes de partitionnement ont beaucoup plus de mal à identifier correctement les classes, d'autant plus que la densité globale du réseau est faible. On s'est intéressé ici à la façon dont le bootstrap clustering réagit face à des problèmes de différentes difficultés. La méthode est appliquée en utilisant la modification de type élongation avec $e = 0.1$, sur des grands réseaux simulés ($v = 6000, n = 100$) avec une densité intra-classe fixé à $d_i = 0.3$, et des densités inter-classe de variables (Fig. 3 page suivante).

1. [9] L'indice de Rand usuel correspond au rapport entre le nombre de paires d'éléments classés de la même façon (réunis ou séparés dans les deux partitions) divisé par le nombre total de paires. L'indice de Rand corrigé prend en compte ce qui est du au hasard (expected by chance), et prend une valeur nulle pour des partitions aléatoires.

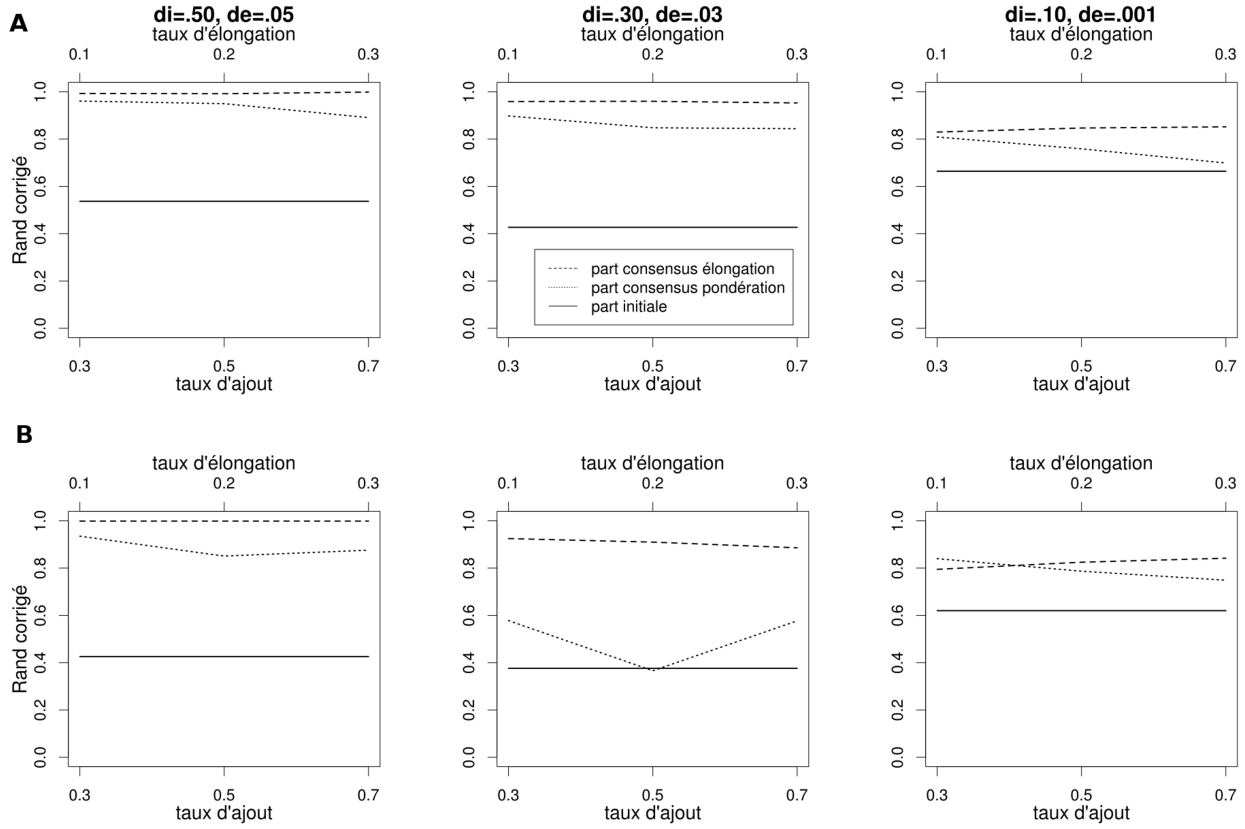



**Figure 2.** Indice de Rand corrigé entre les partition calculés et la partition source, **A** sur des réseaux de taille moyenne avec différentes densités, puis **B** sur des réseaux de grande taille.

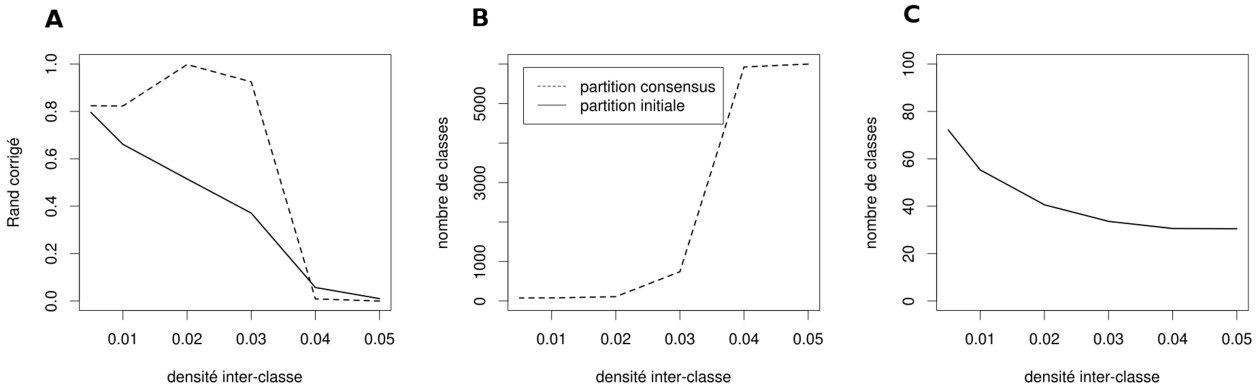



**Figure 3.** Évolution de **A** la proximité des partitions initiale et consensus, **B** du nombre de classes de la partition consensus et **C** du nombre de classes de la partition initiale selon la difficulté du problème.

Lorsque le problème est très facile ($d_e = 0.005$), la partition initiale est déjà très proche de la partition de référence, et la partition consensus apporte peu d'amélioration. Lorsque le problème est très difficile ($d_e = 0.04, 0.05$), les partitions initiale et consensus sont très mauvaises. Par contre, dans des cas intermédiaires ($d_e = 0.01, 0.02, 0.03$), la partition initiale reste assez éloignée de la partition de référence, alors que la partition consensus s'en rapproche fortement, autant en terme d'indice de Rand corrigé que de nombre de classe. Il est intéressant de noter que lorsque le bootstrap clustering s'applique à un réseau sans communautés, il renvoie une partition quasi atomique alors que les méthodes de partitionnement à un seul passage (comme TFit) renvoient un ensemble de classes, certes utilisables mais n'ayant aucun sens.

# 3 Validation biologique

## 3.1 Description des réseaux

| | protéines | interactions | densité |
|---|---|---|---|
| umbilical | 1135 | 2010 | .00312 |
| LCIN | 1902 | 4148 | .00229 |
| HQ12 | 6290 | 18345 | .00092 |
| MQ12 | 14693 | 80931 | .00074 |

**Table 2.** Description des réseaux biologiques étudiés.

Les réseaux étudiés sont de différentes tailles ; le plus petit correspond au réseau entre protéines dont les gènes sont exprimés dans le cordon ombilical [10] (umbilical), le second correspond au Largest Common Interaction Network, le réseau entre protéines dont les gènes sont exprimés ubiquitairement [10] (LCIN), le troisième est un interactome humain de haute qualité (HQ12), et le dernier est un interactome humain plus large (MQ12) (Table 2).

## 3.2 Comparaison de partitions

Les partitions calculées sont comparées avec des références biologiques, qui sont :
- les groupes de protéines annotées au même terme Gene Ontology Biological Process (GO_BP) [11],
- les groupes de gènes GSEA Canonical Pathways, issues de bases de données de voies de signalisation et correspondant aux représentations canoniques des processus biologiques (GSEA) [12],
- les complexes protéiques de mammifères décrits dans la base de données CORUM (complexes) [13].

Chacune de ces références est considérée comme une classification de référence, pour être comparée aux partitions calculées.

La première étape consiste à vérifier que ces partitions soient composées des mêmes éléments, donc de réduire la classification de référence, mais aussi la partition calculée. Ensuite, seuls les groupes de moins de 200 éléments sont utilisés.

La comparaison est basée sur la procédure proposée par [14], qui a pour intérêt de bien gérer les classifications chevauchantes. Cette procédure compare tous les clusters deux à deux en utilisant l'indice de Précision-Rappel pour établir un indice de similarité entre les deux partitions.

## 3.3 Homogénéité fonctionnelle

Pour évaluer la qualité des partitions biologiques, un score d'homogénéité fonctionnelle a été mis en place à partir d'annotations Gene Ontology (GO). Ce score ne mesure pas l'enrichissement d'une fonction dans une classe, mais bien l'homogénéité d'une fonction dans une classe.

### 3.3.1 Annotation fonctionnelle

Les classes sont annotées avec les termes GO biological process (BP), en inférant un terme à une classe s'il est associé à au moins 50 % des protéines de la classe. Cette annotation prend en compte les liens de parenté entre les termes, et seuls les plus précis sont inférés à la classe. Les termes ayant une précision inférieure à 0.3 ne sont pas retenus, donc il est possible que certaines classes ne soient pas annotées. La précision utilisée ici est décrite dans [15] et est une valeur entre 0 et 1, 0 étant associé au terme racine, le moins précis, et 1 aux termes sans enfants des plus longues branches de l'ontologie.

### 3.3.2 Score d'homogénéité fonctionnelle des classes

Pour chaque classe, un score d'homogénéité fonctionnelle est établi. Parmi les termes inférés à une classe, on cherche celui qui est à la fois le plus présent parmi les annotations des protéines et le plus précis pour déterminer le score, selon l'équation suivante :

$$Sc = \max_t \left( \frac{3}{\left(\frac{1}{prc(t)}\right) + \left(\frac{2}{freq(t)}\right)} \right)$$

Le score $Sc$ d'une classe est calculé en maximisant la moyenne harmonique entre $prc(t)$ qui est la précision d'un terme *t*, et $freq(t)$ qui est sa fréquence dans la classe. On a ici choisi de donner plus de poids à la fréquence, qui représente directement l'homogénéité d'un terme dans une classe. Une classe non annotée a un score nul.

### 3.3.3 Score d'homogénéité fonctionnelle de la partition

Pour évaluer l'ensemble de la partition, le score $Sc$ de chaque classe annotée est pris en compte, pondéré par des caractéristiques globales de la partition :

$$Sp = \frac{\sum \left( \frac{5}{\left(\frac{3}{Sc(c)}\right) + \left(\frac{1}{annot}\right) + \left(\frac{1}{size(c)}\right)} \right)}{n}$$

Où $Sp$ est le score de la partition, *n* est le nombre de classes annotées, $annot$ est la fréquence de classes annotées dans la partition, et pour chaque classe *c*, $Sc(c)$ est le score de la classe *c*, et $taille(c)$ est le rapport entre la taille de la classe *c* et la taille moyenne des classes de la partition. Ce score $Sp$ varie entre 0 et 1 ; plus il est proche de 1 et plus la partition est bien annotée, et contient des classes fonctionnellement homogènes.

## 3.4 Application aux réseaux biologiques

La méthode a été appliquée à des réseaux PPI humains décrit précédemment pour déterminer l'amélioration en terme de pertinence biologique apportée par le bootstrap clustering. Les paramètres sont choisis d'après les résultats des simulations, c'est à dire la modification de réseau de type élongation, avec un taux $e = 0.2$. N'ayant plus de partition source à laquelle comparer les partitions calculées, des références biologiques ont été utilisées, tel que les complexes, les groupes fonctionnels GO_BP et les groupes de références GSEA. Étant donné que les classifications de références sont fortement chevauchantes, contrairement aux partitions calculées, l'indice de Rand ne peut plus être utilisé. C'est pourquoi une autre procédure de comparaison entre classifications est utilisée, qui compare tous les clusters d'une partition avec tous les clusters de l'autre partition en utilisant l'indice de Precision-Rappel (Fig. 4 page suivante). Pour chacun des réseaux et pour chacune des références utilisées, la partition consensus est toujours plus proche de la référence que la partition initiale, donc la partition consensus semble avoir une meilleure pertinence biologique que la partition initiale.

## 3.5 Homogénéité fonctionnelle

Une autre façon de mesurer la pertinence biologique d'une partition est d'évaluer l'homogénéité fonctionnelle de chaque classe obtenue. Pour cela, un score basé sur les annotations des protéines formant les classes a été développé. Ce score est calculé sur les partitions des 4 réseaux étudiés, et les scores de partitions initiales et consensus ont été comparés (Fig. 5 page suivante). Dans tous les cas, le score de la partition consensus est meilleur que celui la partition initiale. De plus, le bootstrap clustering calcule des partitions qui permettent l'annotation d'un plus grand pourcentage de classes.

Plus précisément, bien que l'amélioration du score ne soit pas très importante pour les deux petits réseaux, elle est évidente pour les deux interactomes, HQ12 et MQ12. Cette différence de comportement est probablement due au fait que umbilical et LCIN sont des sous-réseaux d'un interactome, et ont donc une structure différente des interactomes plus larges tels que HQ12 et MQ12. Avec umbilical et LCIN, on se trouve probablement dans le cas de problèmes faciles pour l'algorithme, ce qui explique que le bootstrap clustering apporte peu d'amélioration par rapport à la partition initiale. Il est alors intéressant de noter que dans ce cas, les grands interactomes correspondent à des problèmes plus difficiles, mais pour lesquels le bootstrap clustering apporte des solutions pertinentes.

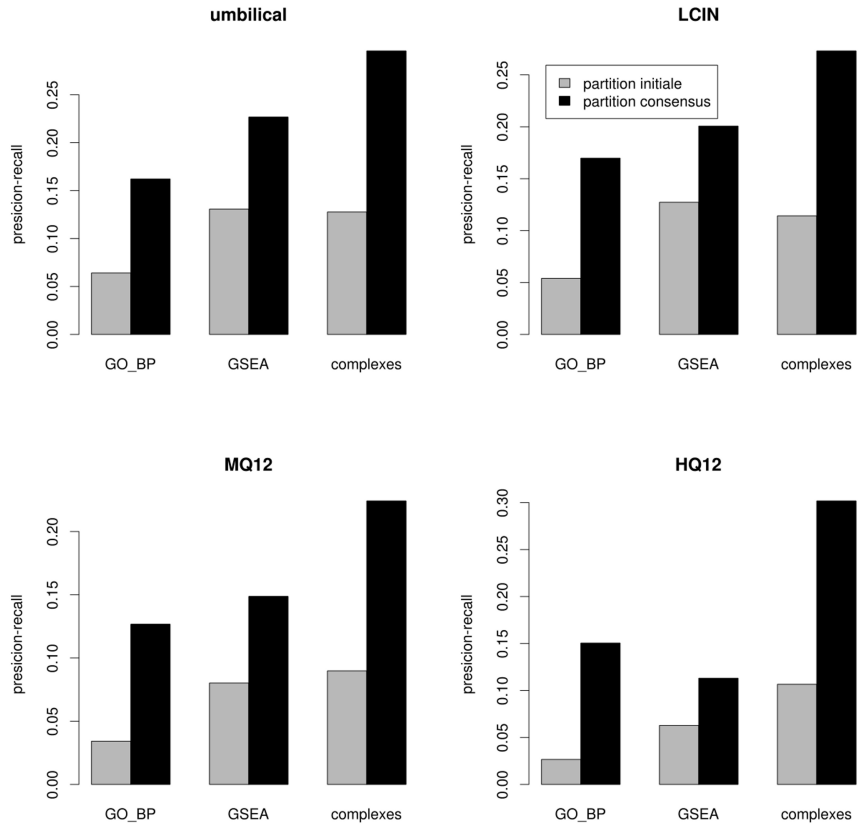



**Figure 4.** Proximité entre les partitions calculés et les références biologiques pour chacun des réseaux étudiés.

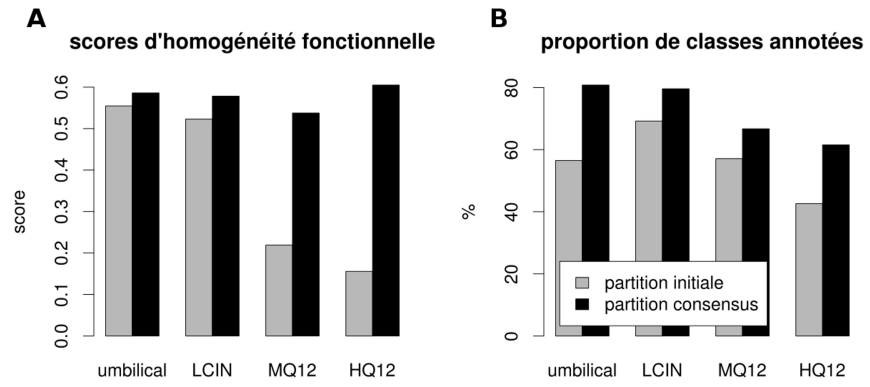



**Figure 5.** **A** Comparaison des scores d'homogénéité fonctionnelle entre les partitions initiales et consensus. **B** Pourcentage de classes annotées dans les partitions initiales et consensus.

## 4 Discussion conclusion

Une des questions posée ici est la pertinence biologique du partitionnement des réseaux PPI. Sachant que ces réseaux sont largement incomplets et bruités, donc quelle est la valeur des résultats biologiques obtenus ? De nombreuses réserves pourraient être émises, mais on observe que ces analyses donnent des résultats biologiquement cohérents, dont certains sont validés expérimentalement [16]. Ces observations montrent que bien qu'incomplets et bruités, les réseaux PPI sont robustes, dans le sens où l'information contenue dans les interactomes représente toujours de façon cohérente l'organisation modulaire des processus biologiques et leurs interactions.

Cependant, le développement d'une méthode de partitionnement nécessite une connaissance *a priori* de la structure du réseau étudié. Pour les réseaux PPI, où seule une fraction du réseau est connu, les caractéristiques topologiques du réseau entier ne peuvent être qu'estimées [17]. C'est une des raisons pour laquelle il est impossible de construire des réseaux simulés similaires aux réseaux PPI, ce qui limite l'évaluation des performances des méthodes sur des simulations. Il est possible d'obtenir des résultats plus robustes en utilisant plusieurs modèles de réseaux simulés, mais il est finalement nécessaire d'appliquer la méthode sur des réseaux biologiques pour vérifier la pertinence fonctionnelle des résultats.

Lorsque la qualité d'une partition biologique doit être évaluée, un moyen simple est de la comparer avec une partition de référence connue. Cependant, cette méthode n'est pas directement applicable car il n'existe pas de partition de référence biologique. Les données s'en rapprochant le plus sont des données orthogonales, telles que des complexes ou des groupes de protéines impliquées dans un même processus biologique que l'on s'attend à retrouver dans un même module fonctionnel. Ces comparaisons sont d'autant plus complexes que ces classifications biologiques sont fortement chevauchantes (donc ne sont pas des partitions au sens strict du terme), ce qui est rarement le cas des partitions calculées. Cette différence structurelle fait que les indices classiques de distance entre partition (Rand, Jaccard) sont peu adaptés à ces comparaisons.

Les interactomes représentent l'ensemble des interactions possibles dans un organisme, sans informations contextuelles. Mais on sait que dans chaque condition cellulaire, les interactions changent. Donc en réalité, il n'y a pas un seul réseau PPI pour un organisme, mais autant de réseaux qu'il y a de conditions cellulaires différentes, et donc autant de partitions. Bien qu'il n'y ait encore que peu d'informations spatio-temporelles nécessaires à des analyses aussi spécialisées, il faut garder à l'esprit que ces interactomes contextualisés correspondent à des réalités biologiques, et qu'il sera nécessaire, pour les analyser, d'avoir aussi de nombreuses références contextualisées.

Lors de l'analyse d'un réseau PPI, se pose souvent la question de quel algorithme utiliser. Pour chaque algorithme différent appliqué à un même réseau, les partitions données peuvent être différentes. Dans ce cas, quelle est la meilleure partition ? Le calcul d'une partition consensus, tel qu'il est utilisé ici, peut être une réponse intéressante, pour combiner les résultats de différents algorithmes. C'est une piste qui reste à explorer, mais dans laquelle le choix des algorithmes utilisés va être déterminant, notamment selon les structures des partitions calculées (chevauchantes ou non, totalement recouvrantes ou non). De plus, sachant que de nombreuses protéines peuvent être impliquées dans plusieurs fonctions, on s'attend à ce qu'elle appartiennent à plusieurs classes. Le calcul d'une partition consensus pourrait permettre de transformer un profil de partitions strict en un système de classes chevauchantes plus représentatif de la réalité biologique.

La méthode du boostrap clustering cherche à tenir compte de ces différentes problématiques pour proposer une partition ayant la meilleure qualité biologique possible. On peut observer que paradoxalement, l'ajout répété d'un bruit raisonnable dans un réseau par ailleurs incomplet et bruité permet d'améliorer la partition obtenue. Cette étude mène donc à reconsidérer l'impact de l'incertitude dans les réseaux PPI. Cette incertitude, ou bruit (faux positifs et faux négatifs), si longtemps considérée comme une forte limitation des données d'interactions protéine-protéine, pourrait être dépassée et améliorer le résultat d'un partitionnement d'un réseau biologique lorsqu'une méthode de bootstrap clustering est appliquée.

Enfin, de manière intéressante, il est à noter que lors de l'analyse des réseaux biologiques, la différence entre partition initiale et partition consensus semble correspondre à l'influence qu'a le bruit dans les données sur le partitionnement. Bien que des analyses supplémentaires soient nécessaires pour le confirmer, la méthode de bootstrap clustering pourrait permettre de quantifier le bruit des données biologiques.

## Remerciements

Nous remercions Anaïs Baudot et Charles Chapple pour de fructueuses discussions. Ce travail est financé par une bourse doctorale de la fondation Axa pour la Recherche attribuée à B.R.

## Références


[1] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, "A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae," *Nature*, vol. 403, pp. 623–627, Feb. 2000.

[2] S. Fields and O.-k. Song, "A novel genetic system to detect protein-protein interactions," *, Published online : 20 July 1989 ; | doi :10.1038/340245a0*, vol. 340, pp. 245–246, July 1989.

[3] A. J. M. Walhout, S. J. Boulton, and M. Vidal, "Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm," *Yeast*, vol. 17, pp. 88–94, June 2000.

[4] K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A.-S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A.-L. Barabasi, and M. Vidal, "An empirical framework for binary interactome mapping," *Nat Meth*, vol. 6, pp. 83–90, Jan. 2009.

[5] T. Aittokallio and B. Schwikowski, "Graph-based methods for analysing networks in cell biology," *Briefings in Bioinformatics*, vol. 7, pp. 243–255, Jan. 2006.

[6] P. Gambette and A. Guénoche, "Bootstrap clustering for graph partitioning," *RAIRO - Operations Research*, vol. 45, pp. 339–352, Mar. 2012.

[7] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, p. 297, July 1945.

[8] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 8577 –8582, June 2006.

[9] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, Dec. 1985.

[10] O. Souiai, E. Becker, C. Prieto, A. Benkahla, J. De Las Rivas, and C. Brun, "Functional integrative levels in the human interactome recapitulate organ organization," *PLoS ONE*, vol. 6, p. e22051, July 2011.

[11] 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," *Nature genetics*, vol. 25, pp. 25–29, May 2000. PMID : 10802651.

[12] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 15545–15550, Oct. 2005.

[13] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and H.-W. Mewes, "CORUM the comprehensive resource of mammalian protein complexes–2009," *Nucleic Acids Research*, vol. 38, pp. D497–D501, Nov. 2009.

[14] S. Brohee and J. van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, no. 1, p. 488, 2006.

[15] C. Herrmann, S. Bérard, and L. Tichit, "SimCT a generic tool to visualize ontology-based relationships for biological objects," *Bioinformatics*, vol. 25, pp. 3197 –3198, Dec. 2009.

[16] M. A. Pujana, J.-D. J. Han, L. M. Starita, K. N. Stevens, M. Tewari, J. S. Ahn, G. Rennert, V. Moreno, T. Kirchhoff, B. Gold, V. Assmann, W. M. ElShamy, J.-F. Rual, D. Levine, L. S. Rozek, R. S. Gelman, K. C. Gunsalus, R. A. Greenberg, B. Sobhian, N. Bertin, K. Venkatesan, N. Ayivi-Guedehoussou, X. Solé, P. Hernández, C. Lázaro, K. L. Nathanson, B. L. Weber, M. E. Cusick, D. E. Hill, K. Offit, D. M. Livingston, S. B. Gruber, J. D. Parvin, and M. Vidal, "Network modeling links breast cancer susceptibility and centrosome dysfunction," *Nature Genetics*, vol. 39, no. 11, pp. 1338–1349, 2007.

[17] A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Scientific American*, vol. 288, pp. 60–69, May 2003. PMID : 12701331.

# Knowledge-based zooming for metabolic models

Anna ZHUKOVA[1] and David James SHERMAN[1]

INRIA / Université Bordeaux 1 / CNRS joint project-team MAGNOME, 351, cours de la Libération, F-33405 Talence, Cedex, France

{anna.zhukova, david.sherman}@inria.fr

**Keywords** metabolic modelling, knowledge databases, genome-scale.

Genome-scale metabolic models for new organisms include thousands of reactions. In most cases these reactions are automatically inferred by methods that combine databases of reactions and pathways with genomic information and existing models for similar organisms [1]. Genomic data for the new organism is compared to the data of the reference organism, to find genomic evidence such as the presence of catalysing enzymes for the reactions conserved in the new organism. Starting from the inference of a draft model, the model refinement process includes several iterations of model analysis, error detection, and improvement [2]. The models produced at each iteration are intended for computer simulation, and so describe all the reactions thought to participate in the organism's metabolism. Although automatic model inference tools and genome comparison methods are becoming more and more advanced, they still may leave gaps in the model or add erroneous reactions. Thus, model evaluation by human experts remains important at all the iteration steps. However, because of their completeness, genome-scale models are too detailed and complicated to be easily understood by a human. The abundance of reactions in the model may hide errors.

For example, if in a genome-scale model of an yeast *Yarrowia lypolitica* (*MODEL1111190000* [3]) the enzyme *EC 2.3.1.16* were missing, the whole group of *Acyl-CoA:acetyl-CoA C-acyltransferase* reactions participating in the *Beta-oxidation of fatty acids* pathway [4] would be eliminated: one for each of the six *3-oxoacyl-CoA* species (*3-oxodecanoyl-CoA*, *3-oxohexacosanoyl-CoA*, *3-oxolauroyl-CoA*, *3-oxooctadecanoyl-CoA*, *3-oxopalmitoyl-CoA*, and *3-oxotetradecanoyl-CoA*) present in the model. However, the absence of these six reactions would be hidden by the other 59 reactions in the constitutive peroxisome of *Yarrowia lypolitica*, and a human expert may have difficulty noticing the error.

To aid human understanding of these complete models, we developed a method for knowledge-based zooming that provides a higher-level view of a model while keeping its essential structure. The zooming process groups chemical species present in the model into semantically equivalent classes, and merges them into a generalized chemical species. The *ubiquitous species*, that participate in many reactions and are common to most of the models, e.g. *water*, *ATP*, *oxygen*, do not need to be generalized: Each of them forms a trivial equivalence class. The other species are divided into non-trivial equivalence classes, based on their hierarchical relationships in the *ChEBI* ontology [5], and generalized accordingly. For example, *3-oxodecanoyl-CoA*, *3-oxohexacosanoyl-CoA*, and *3-oxolauroyl-CoA* can be all generalized into *3-oxoacyl-CoA*. Reactions that involve same generalized chemical species are then factored together into a generalized reaction. The zooming process is represented in figure Fig. 1.

By applying this process, we can build a simplified model that focusses on the high level relationships. Our method obeys several consistency restrictions, such as conserving the number of distinct species participating in each reaction (i.e. preserving reaction stoichiometry); and preserving connectivity, i.e. for every pair of reactions sharing a reactant/product in the initial model, the "zoomed out" reactions share the "zoomed out" reactant/product. We implemented our method (in Python) and applied it to several genome-scale metabolic models.

## Acknowledgements

Anna Zhukova was supported by a CORDI-S doctoral fellowship from Inria.

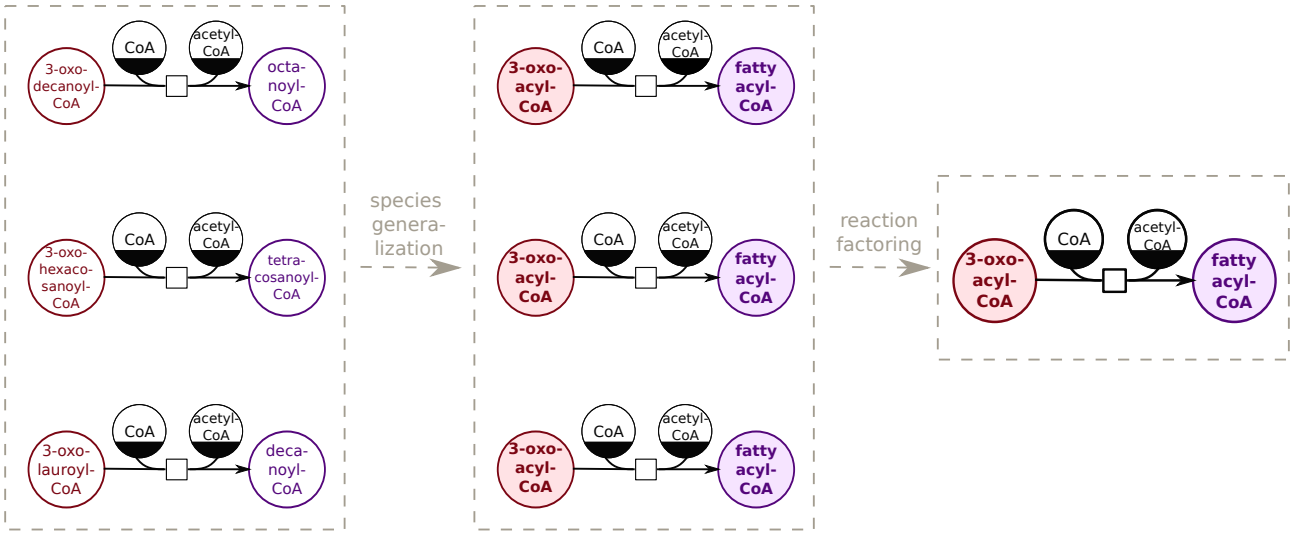



**Figure 1.** Model zooming process.

## References

[1] N. Swainston, K. Smallbone, P. Mendes, D. B. Kell and N. W. Paton, The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *Journal of integrative bioinformatics*, 8, 186 (2011).

[2] I. Thiele and B. Ø. Palsson, A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5, 93–121 (2010).

[3] N. Loira, T. Dulermo, J.-M. Nicaud and D. J. Sherman, A genome-scale metabolic model of the lipid-accumulating yeast Yarrowia lipolytica. *BMC Systems Biology*, 6, 35 (2012).

[4] D.E. Metzler, Biochemistry: The Chemical Reactions of Living Cells. *Elsevier Science*, 2001.

[5] P. de Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner and C. Steinbeck, Chemical Entities of Biological Interest: an update. *Nucleic Acids Research*, 38, D249–D254 (2010).

# Session 7: Réseaux biologiques

## Conférence invitée

PEDRO MENDES

School of Computer Science and Manchester Institute of Biotechnology,
University of Manchester & Virginia Bioinformatics Institute, Virginia Tech

### From Genomes to Large-Scale Kinetic Models of Metabolism

Advances in genomic sequencing have resulted in a large number of complete genome sequences, which are valuable to many fields of biology, including metabolism and metabolic modelling. A common practice in genomics is to use these sequences and their annotation to reconstruct the metabolism of a whole organism. Many metabolic reconstructions exist, including some created by community projects that resulted in high-quality consensus networks. Metabolic reconstructions are useful on their own as a reference, but they are even more useful as the basis to build quantitative models. The traditional application has been to characterise optimal metabolic states by flux balance analysis. A more ambitious objective is to create full kinetic models, that are also able to describe the dynamics of the network in computer simulations. A major challenge is the determination of equilibrium constants, rate laws and values of the kinetic constants. The result of this approach produces large-scale kinetic models of metabolism which should be seen as hypotheses about the dynamics of the metabolic network. Those models have low predictive power at first, but are important starting points for further improvement. Our strategy for improving these models is to incorporate parameter values determined in enzyme kinetics experiments. Rounds of modelling and experimentation are proposed to identify the enzymes that should be studied experimentally and that best improve the accuracy of the large-scale model. We propose that these large-scale kinetic models should become a major focus of the genomics community as a means to use the vast amounts of sequence information into biochemical knowledge.

# Quantitative comparison of one-step and two-step models of gene expression

Valentin Zulkower[1], Jean-Luc Gouzé[2] and Hidde de Jong[1]

[1] Institut National de Recherche en Informatique et en Automatique (INRIA), Unité de recherche Rhône-Alpes, 655 avenue de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France.
{valentin.zulkower, hidde.de-jong}@inria.fr

[2] Institut National de Recherche en Informatique et en Automatique (INRIA), Unité de recherche Sofia Antipolis, BP 93, 06902 Sophia-Antipolis Cedex, France.
jean-luc.gouze@inria.fr

**Abstract** *When modeling gene expression, transcription and translation can be lumped into a single step in order to simplify analysis and simulation. While this simplification is widely used, its consequences on the model predictions are rarely evaluated. In this article we study how biological parameters and changes in promoter activity influence the relative difference between the outputs of the one-step and two-step models, and how reporter genes can be used to estimate this difference from experimentally measured promoter activity profiles. The application of this method to two genes of E. coli shows that the difference will be typically of the order of 1%, with possible transient increases in case of rapid regulation changes.*

**Keywords** Systems biology, model reduction, gene expression model, fluorescent reporter genes.

## Comparaison quantitative des modèles d'expression génique à une et deux étapes.

**Résumé** *Lors de la modélisation de l'expression d'un gène, les étapes de transcription et traduction peuvent être regroupées en une seule étape afin de simplifier l'analyse et la simulation du modèle. Bien que cette simplification soit courament utilisée, ses consequences sur les prédictions du modèle sont rarement évaluées. Dans cet article nous étudions comment la différence relative entre les prédictions des modèles à une et deux étapes est influencée par les paramètres biologiques et les changements de l'activité du promoteur, et comment des gènes rapporteurs peuvent être utilisés pour estimer cette différence à partir de profils d'activité génique mesurés expérimentalement. L'application de cette méthode à l'étude de deux gènes de l'entero-bacterie E. coli montre que cette difference sera typiquement de l'ordre de 1%, avec potentiellement des pics transitoires en cas de rapides changements de la régulation.*

**Mots-clés** Biologie des systèmes, réduction de modèles, gènes rapporteurs fluorescents.

## 1 Introduction

Most models of gene expression do not capture the full complexity of protein synthesis, but rather distinguish two prime steps, transcription and translation, denoting respectively the synthesis of messenger RNA (mRNA), and the production of proteins from the information contained in the mRNA. Such two-step gene expression models are usually the building blocks of larger models describing the network of interactions of several genes, mRNAs, proteins, and metabolites. To make the network simpler to analyze and easier to handle computationally, it may be worthwhile to simplify even further this two-step model, and lump the entire gene expression process into a single step.

We will focus on Ordinary Differential Equations (ODE) models, for which both two-step or one-step models of gene expression can be used. In this context, the reduction of two-step models to one-step models is based on the assumption that the mRNA concentrations are in quasi-steady state, in the sense that they adapt almost instantaneously to changes in the promoter activity. This assumption, which makes it possible to overlook the variations of the mRNA concentration, and write variations of the protein concentration directly as

a function of the promoter activity, is known as the quasi-steady-state approximation (QSSA). Simplifications based on similar arguments have been extensively studied in enzyme kinetics, notably in the context of the reduction of mass-action models to the Michaelis-Menten rate law [1,2]. Gene expression models have been less studied from this point of view. Under which conditions is it justified to simplify two-step models to one-step models, and to which extent will this simplification influence the model predictions ?

In this extended abstract of a paper submitted for publication [3] we study the difference over time of the predictions of the one-step and the two-step model, which we will refer to as the model reduction error. We will first see that writing the model in terms of total amounts of molecules instead of concentrations of molecules can provide insight on the relevant parameters and variables for the study of this error. We will then show how the relevant variables can be experimentally estimated by means of fluorescent reporter experiments. As an application we measure the model reduction error in the case of the gene *crp*, whose dynamics are observed during a typical growth-phase transition. Our results show that, for physiologically-relevant parameter sets, the relative error induced by the model reduction will remain below the observed experimental variability.

## 2 One-step and two-step models of gene expression

Let $m(t)$ [$\mu$M] and $p(t)$ [$\mu$M] denote the time-varying concentrations of mRNA and protein, respectively, with time $t \in \mathbb{R}_+$. The two-step model is a system of two ODEs for $m(t)$ and $p(t)$, describing transcription and translation respectively:

$$\frac{d}{dt}m(t) = \kappa_m\, f(t) - (\gamma_m + \mu(t))\, m(t), \quad m(0) = m_0, \tag{1a}$$

$$\frac{d}{dt}p(t) = \kappa_p\, m(t) - (\gamma_p + \mu(t))\, p(t), \quad p(0) = p_0, \tag{1b}$$

with $\kappa_m$ [M min$^{-1}$] the maximum synthesis rate of mRNA, and $\kappa_p$ [min$^{-1}$] the maximum synthesis rate of protein per unit mRNA. The function $f(t) : \mathbb{R}_+ \to [0, 1]$ describes the modulation over time of the rate of mRNA synthesis by transcriptional regulators. mRNA and protein are degraded in a first-order reaction with degradation constants $\gamma_m$ [min$^{-1}$] and $\gamma_p$ [min$^{-1}$], respectively, and diluted through the growth of the cell population, with growth rate $\mu(t)$[min$^{-1}$]. The degradation constants are related to the half-lives of mRNA and protein, denoted by $\tau_{m,1/2}$ and $\tau_{p,1/2}$ respectively, as follows: $\tau_{m,1/2} = \ln 2/\gamma_m$ and $\tau_{p,1/2} = \ln 2/\gamma_p$. The growth rate of the cell population can be written

$$\mu(t) = \frac{1}{V(t)}\frac{d}{dt}V(t)$$

where $V(t)$ [L] denotes the volume of the population.

The typical half-life of mRNA in bacteria (on the order of a few minutes [4]) is small compared to the time-scale of other phenomena like cell division (from tens of minutes in rich media to hours in minimal media [5,6]) or protein degradation (hours for almost all proteins [7,8]). We thus conclude that $\gamma_m \gg \gamma_p$ and $\gamma_m \gg \mu$, which implies that mRNA concentrations reach their steady state much faster than protein concentrations. It motivates the usual assumption that the mRNA concentration is always at quasi-steady, translated in terms of equations as $dm(t)/dt = 0$ , from which follows that

$$m(t) = \kappa_m f(t)/\big(\gamma_m + \mu(t)\big).$$

In this expression we can use the above-mentioned fact that the growth rate is small compared to the degradation constant of the mRNA, and write $m(t) = \kappa_m\, f(t)/\gamma_m$. Re-injecting this equality into the first equation leads to a simplified (and approximate) system of a single ODE describing the evolution of the protein concentration with the same regulatory input $f(t)$

$$\frac{d}{dt}\hat{p}(t) = \frac{\kappa_m\, \kappa_p}{\gamma_m}\, f(t) - (\gamma_p + \mu(t))\, \hat{p}(t), \quad \hat{p}(0) = \hat{p}_0. \tag{2}$$

In Eq. 2 the transcription and translation processes are lumped into one step, and the ratio $\kappa_m\kappa_p/\gamma_m$ can be treated as a single phenomenological synthesis parameter, thus significantly reducing the number of parameters of the model. Moreover, as the *fast* variable $m(t)$ is no longer explicitly considered, this model will be easier to solve numerically (reduction of stiffness). The one-step model of gene expression is schematically compared with the two-step model in Fig. 1a-b.

A mathematical basis generally invoked for the application of the QSSA is Tykhonov's theorem for dynamical systems [9,10]. However, this theorem only gives a limit behavior of the system when some scaling parameter converges towards infinity. Moreover, it usually does not consider input variables, which vary on a particular time-scale themselves. How well does the one-step model approximate the two-step model in a particular context, as defined by specific half-lives of mRNA and protein, specific initial conditions, and a specific promoter activity and population growth rate? In our study we will focus on the relative model reduction error, which provides a measure of the quality of the approximation:

$$\Delta(t) = \frac{\mid p(t) - \hat{p}(t) \mid}{p(t)}. \tag{3}$$

In the next section we will show how this error can be estimated through reporter gene experiments.

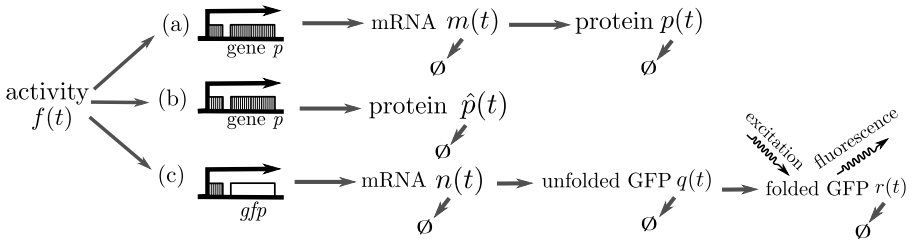



**Figure 1.** Two-step (a) and one-step (b) models of gene expression, and reporter gene expression model (c).

## 3 Estimation of the model reduction error through reporter gene experiments

The system described in Eq. 1 undergoes two distinct external perturbations, $f(t)$ and $\mu(t)$, reflecting the influences of transcriptional regulation and changes in the cell population volume, respectively. It is possible to aggregate these effects by introducing the following variables:

$$M(t) = V(t)\,m(t), \quad P(t) = V(t)\,p(t), \quad F(t) = V(t)\,f(t), \quad \hat{P}(t) = V(t)\,\hat{p}(t). \tag{4}$$

$M(t)$, $P(t)$, and $\hat{P}(t)$, expressed in mole units, represent the amounts of mRNA and protein summed over the volume of the cell population, while $F(t)$ is the the cumulative activity of all promoters of a specific gene in the cell population, and has the unit mole $\text{min}^{-1}$. This change of variables allows the two-step model to be rewritten as follows:

$$\frac{d}{dt}M(t) = \kappa_m\,F(t) - \gamma_m\,M(t), \tag{5a}$$

$$\frac{d}{dt}P(t) = \kappa_p\,M(t) - \gamma_p\,P(t). \tag{5b}$$

while the reduced model becomes

$$\frac{d}{dt}\hat{P}(t) = \frac{\kappa_m\kappa_p}{\gamma_m}\,F(t) - \gamma_p\,\hat{P}(t). \tag{6}$$

Notice that the growth-dilution terms have disappeared from the reformulated models, as we do no longer consider concentrations but total amounts of molecules in a (possibly) expanding volume. In the terminology of physics, the reformulation of the model implies a change from intensive to extensive variables.

Conveniently, the model reduction error defined in Eq. 3 can be written as a function of the extensive variables as well:

$$\Delta(t) = \frac{\mid p(t) - \hat{p}(t) \mid}{p(t)} = \frac{\mid V(t)p(t) - V(t)\hat{p}(t) \mid}{V(t)p(t)} = \frac{\mid P(t) - \hat{P}(t) \mid}{P(t)}. \tag{7}$$

In what follows, we study $\Delta(t)$ by means of the new systems given by Eq. 5 and Eq. 6.

Notice that in this reformulation the relative error $\Delta(t)$ does not depend on $f(t)$ and $\mu(t)$ separately, but is affected by their joint influence on the variable $F(t)$. More precisely, Eq. 5 and Eq. 6 describe two linear filters with input $F(t)$ and outputs $P(t)$ and $\hat{P}(t)$ respectively. It is then possible to show that the relative model reduction error $\Delta(t)$ will be insensible to the initial conditions and to the values of the production rates $\kappa_m, \kappa_p$, and that it will remain unchanged if the input $F(t)$ is replaced by a proportional input $F_e(t) = \alpha F(t)$, $\alpha > 0$ . This last point comes with interesting practical implications, as we will show that such proportional profile $F_e$ can be easily estimated from population-level fluorescence signals measured in reporter gene experiments.

Current reporter gene technologies, based on Green Fluorescent Proteins (GFPs) provide an excellent means to measure promoter activities *in vivo* and in real time ([11,12]). The underlying principle of the technology is to fuse the promoter region of a gene of interest to a reporter gene. The expression of the reporter gene generates a visible fluorescence signal $I(t)$ from which the promoter activity of the reporter gene can be reconstructed (see [13,14] and references therein). Since the promoter regions of the reporter gene and the gene of interest are the same, it is natural to assume that the reconstructed promoter activity is also the promoter activity of the gene of interest.

In earlier work [13], we developed and validated a measurement model for the interpretation of fluorescence data. This model is a variant of the two-step model, taking into account that the fluorescent activity of GFP in response to light excitation depends on post-translational modifications, notably the folding of the protein to an appropriate conformation, including the autocatalytic formation of the chromophore [12]. This maturation process gives rise to an additional reaction step from inactive to active GFP. As a consequence, the state variables of the measurement model are *gfp* mRNA ($n(t)$ [$\mu$M]), inactive GFP ($q(t)$ [$\mu$M]), and active GFP ($r(t)$ [$\mu$M]), as represented in Fig. 1c . The expression of the fluorescent protein can be modeled as follows :

$$\frac{d}{dt}n(t) = \kappa_n\, f(t) - (\gamma_n + \mu(t))\, n(t), \tag{8a}$$

$$\frac{d}{dt}q(t) = \kappa_q\, n(t) - (\gamma_q + \kappa_r + \mu(t))\, q(t), \tag{8b}$$

$$\frac{d}{dt}r(t) = \kappa_r\, q(t) - (\gamma_r + \mu(t))\, r(t). \tag{8c}$$

The synthesis parameters $\kappa_n$ and $\kappa_q$ correspond to $\kappa_m$ and $\kappa_p$ in the gene expression model of Eq. 1, respectively, while $\kappa_l$ [min$^{-1}$] is the GFP folding constant. The degradation constants of reporter mRNA and of the folded and unfolded reporter proteins are denoted by $\gamma_n$, $\gamma_r$, and $\gamma_q$, respectively. While the degradation constants of the folded and unfolded GFP can be supposed identical ($\gamma_q = \gamma_r$), the degradation constants of the reporter mRNA and reporter protein are different from the degradation constants for the products of the gene of interest, that is, $\gamma_q \neq \gamma_p$ and $\gamma_n \neq \gamma_m$.

The system of Eq. 8 can also be written in terms of extensive variables: the input of the system is then $F(t)$, and the variable $r(t)$ would then be replaced by $R(t)$, denoting the total amount of folded reporter in the medium. The variable $R(t)$ can be assumed proportional to the measured fluorescence of the GFP, $I(t)$. Under this assumption it is possible, by collapsing the system of Eq. 8, to write

$$F(t) \propto F_e(t) = I(t) + a\frac{d}{dt}I(t) + b\frac{d^2}{dt^2}I(t) + c\frac{d^3}{dt^3}I(t), \tag{9}$$

where

$$a = \frac{1}{\gamma_n} + \frac{1}{\gamma_r} + \frac{1}{\gamma_r + \kappa_r}, \qquad b = \frac{1}{\gamma_n \gamma_r} + \frac{1}{\gamma_n(\gamma_r + \kappa_r)} + \frac{1}{\gamma_r(\gamma_r + \kappa_r)}, \qquad c = \frac{1}{\gamma_r \gamma_n(\gamma_r + \kappa_r)}. \quad (10)$$

and were the proportionality factor depends on the variables $\kappa_n, \kappa_q, \kappa_r$.

By injecting the profile $F_e(t)$ into Eq. 5 and Eq. 6 we obtain two profiles, $P(t)$ and $\hat{P}(t)$ respectively, for the protein of interest, and the model reduction error $\Delta(t)$ can be computed using Eq. 7. This way we have estimated the relative difference between the concentration $p(t)$ and its approximated version $\hat{p}(t)$, based on actual promoter activity and growth rate profiles $f(t)$ and $\mu(t)$, but without having to estimate $f(t)$ and $\mu(t)$.

## 4 Application to the study of the gene expression model in *E. coli*

Microorganisms like the enterobacterium *E. coli* use glucose and other carbon sources for growth. *E. coli* has intricate regulatory mechanisms, on both the metabolic and genetic level, to adapt the functioning of carbon metabolism to the availability of different carbon sources in the environment. For instance, when a preferred (rich) carbon source like glucose is depleted, it continues its growth on less preferred (poorer) carbon sources like acetate. A change in carbon source is accompanied by a profound reorganization of metabolic fluxes and of the expression of the genes encoding enzymes of the metabolic reactions [15].

We consider here the gene *crp*, whose expression patterns is typical from those encountered during growth transitions. The gene *crp* encodes the transcription factor Crp that is a pleiotropic regulator of the cell [16]. The complex Crp·cAMP regulates the transcription of hundreds of genes in *E. coli*, many of them enzymes catalyzing reaction steps in carbon metabolism [17,18]. The protein concentration has been shown to vary little [19,20] during the transition phase.

Batch cultures of *E. coli* were grown in a microplate at 37 °C in M9 minimal medium supplemented with glucose. To measure the expression of the genes of interest, we used strains transformed with reporter plasmids carrying a transcriptional fusion of a *gfp* reporter gene and the promoter region of *crp*, respectively [21]. By means of an automated microplate reader, we measured the absorbance $A(t)$ of the culture and the emitted fluorescence $I(t)$ (Fig. 2a). The absorbance is not strictly necessary for our study, as we have shown that the variable of interest, $\Delta(t)$, can be reconstructed from $I(t)$ alone. However, the absorbance will be used as a measure for the population volume, in order to compute the variables $p(t)$, $\hat{p}(t)$, and $m(t)$, and to synchronize datasets from different bacterial colonies.

We applied the method described in Section 3 to infer a profile $F_e(t)$ proportional to the cumulative promoter activity (Fig. 2b), reconstruct the profile of the total amounts of protein $P(t)$ and $\hat{P}(t)$ (Fig. 2c), and compute $\Delta(t)$ (Fig. 2d). For illustration purpose, we also provide (rough) estimates profiles of $f(t)$, $p(t)$, and $\hat{p}(t)$ for each gene, obtained by dividing the global signals $F(t)$, $P(t)$ and $\hat{P}(t)$ by the absorbance. The computation of the promoter activities of *crp* requires the degradation and folding constants of the reporter system to be known. The GFP used in our study was shown to have a half-life of about 17 h in our conditions ($\gamma_q = 0.0007$ min$^{-1}$) and a folding rate $\kappa_r$=0.3 min$^{-1}$ [21]. We took half-lives previously measured in our conditions for the gene *fis* as reference values for $\gamma_p$ and $\gamma_m$ (about 100 min and 1.2 min, respectively, corresponding to $\gamma_p = 0.0065$ min$^{-1}$ and $\gamma_m = 0.56$ min$^{-1}$).

Upon growth arrest, the promoter activity $f(t)$ of *crp* starts to decrease and reaches a two-fold lower stationary level. The relative error stays below 1%, and the profiles of $p(t)$ and $\hat{p}(t)$ overlap to the point of being indistinguishable. In particular, the variance between the curves computed from several replicates (shaded regions in Fig. 2), is much more important than the model reduction error. We conclude that the numerical error induced by the two-step to one-step model simplification is not be considered a critical issue when modeling the expression of *crp* in this experiment.

## 5 Discussion

We have shown that the expression of the gene *crp E. coli* can be modeled using either one-step or two-step models of gene expressions, with only minor quantitative differences (1%) between the two models, as it would

have been for any other gene with slow promoter activity dynamics. Reporter genes experiments can provide the relevant variables for the estimation of the model reduction error. For genes with different dynamics, like *acs*, for which the promoter activity can change several folds over a few minutes under the influence of fast-varying metabolite concentrations, the error $\Delta(t)$ can pike temporarily (to about 5%) in the tens of minutes following the disruption (data not shown, see [3]).

We have presented in this paper a way of systematically checking for the validity of the model reduction using an experimental approach. A more thorough mathematical analysis of the models, reported in [3], can also provide more insight on the dependency between $\Delta(t)$, $F(t)$, and the degradation rates $\gamma_m$ and $\gamma_p$. We can for instance show that the maximum value of the quantity $|(1/\gamma_m)(1/F)dF/dt|$ up to time $t$ is, in many cases, a fine upper bound for $\Delta(t)$ on the same interval. This provides a simple rule of thumbs to check *a priori* the validity of the model reduction. This also hints that rapid changes in the promoter activity will tend to increase the model reduction error, but as discussed above, even in a very disadvantageous case, the error is bounded and transient.

Note that the two-step model used as a starting point in our analysis is itself obtained by simplifying a more detailed model that takes into account the individual reaction steps of transcription and translation (see [22,23]). The reduction of the gene expression model is particularly beneficial to reduce the computation time and numbers of parameters when dealing with models of large gene networks. In this case, much attention must be given to the numerical propagation of the model reduction error. It is possible to show that, for gene cascades where the regulation dependencies have linear or Michaelis-Menten forms, or any composition of such, then the error resulting from the model reduction of one gene will not be amplified through the cascade, and, the simultaneous reduction of the expression model of several genes of the cascade will have a quasi-additive effect on the error on the genes downstream.

As a conclusion, transcription appears to be a transparent step with respect to the general profile of the protein concentrations. For some eukaryotic organisms it has been shown that the delay between transcription and translation has much influence on the modulation of gene expression, and can play a functional role, notably in circadian rhythms [24]. The transcription step may have no such role in bacteria, but can play other roles, like accelerating the gene response (as multiple copies of mRNA can be translated in parallel), and, in some cases, enabling the fine-tuning of gene expression by post-transcriptional regulations through riboswitches or modulations of the mRNA half-life.

## 6 Acknowledgments

This work was supported by the Agence Nationale de Recherche under project GeMCo (ANR-2010-BLAN-0201-02) and INRIA/INSERM under project Colage.

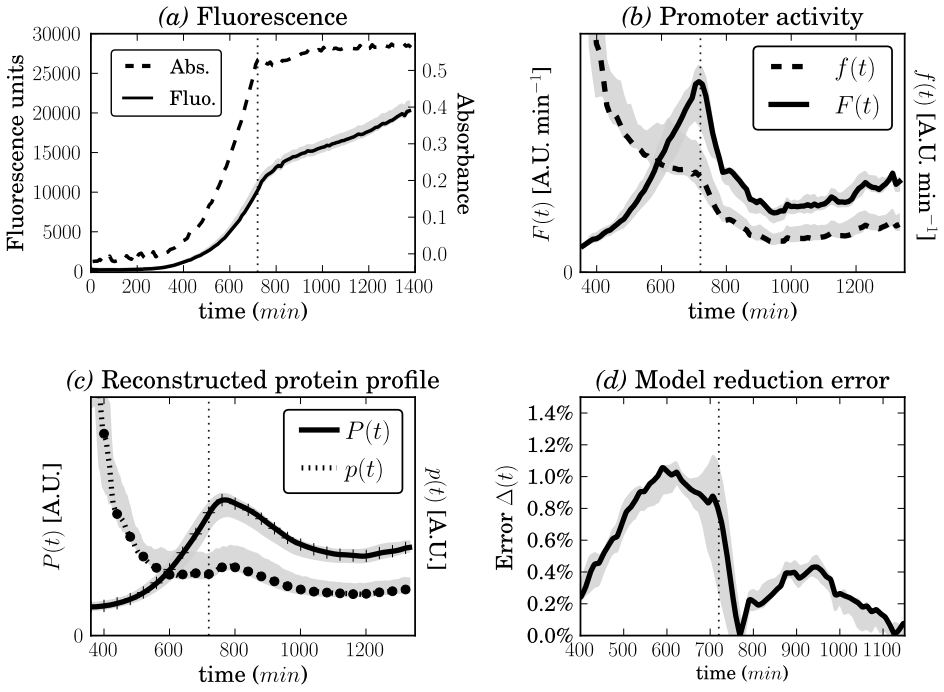



**Figure 2.** Estimation of the model reduction error from reporter experiment data, for the gene *crp*. The wide lines profiles were computed using data from the same well. Shaded areas represent two standard deviations to the mean of at least five replicate wells. Vertical dotted line mark the entry into stationnary phase. In graph *a*, the fluorescence signal has been separated from background noise and the autofluorescence of the cell population, measured on control wells. In graph *c*, symbols $+$ and $\bullet$ mark the values of $\hat{P}(t)$ and $\hat{p}(t)$, computed using the reduced model.

## References


[1] J.A. Borghans, R.J. de Boer, and L.A. Segel. Extending the quasi-steady state approximation by changing variables. *Bulletin of Mathematical Biology*, 58(1):43–63, 1996.

[2] W. Chen, M. Niepel, and P.K. Sorger. Classic and contemporary approaches to modeling biochemical reactions. *Genes and Development*, 24(17):1861–1876, 2010.

[3] V. Zulkower et al. One-step and two-step models of gene expression in bacteria. *submitted for publication*, 2013.

[4] J.A. Bernstein, A.B. Khodursky, P.-H. Lin, S. Lin-Chao, and S.N. Cohen. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proceedings of the National Academy of Sciences of the USA*, 99(15):9697–9702, 2002.

[5] M. Schaechter, O. Maaløe, and N.O. Kjeldgaard. Dependency on medium and temperature of cell size and chemical composition during balanced grown of *Salmonella typhimurium*. *Journal of General Microbiology*, 19(3):592–606, 1958.

[6] K.B. Andersen and K. von Meyenburg. Are growth rates of escherichia coli in batch cultures limited by respiration? *Journal of Bacteriology*, 144(1):114–123, 1980.

[7] R.D. Mosteller, R.V. Goldstein, and K.R. Nishimoto. Metabolism of individual proteins in exponentially growing *Escherichia coli*. *Journal of Biological Chemistry*, 255(6):2524–2532, 1980.

[8] K.L. Larrabee, J.O. Phillips, G.J. Williams, and A.R. Larrabee. The relative rates of protein synthesis and degradation in a growing culture of *Escherichia coli*. *Journal of Biological Chemistry*, 255(9):4125–4130, 1980.

[9] R. Heinrich and S. Schuster. *The Regulation of Cellular Systems*. Chapman & Hall, New York, 1996.

[10] H.K. Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 3rd ed. edition, 2001.

[11] B.N. Giepmans, S.R. Adams, M.H. Ellisman, and R.Y. Tsien. The fluorescent toolbox for assessing protein location and function. *Science*, 312(5771):217–224, 2006.

[12] R.Y. Tsien. The green fluorescent protein. *Annual Review of Biochemistry*, 67:509–544, 1998.

[13] H. de Jong, C. Ranquet, D. Ropers, C. Pinel, and J. Geiselmann. Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. *BMC Systems Biology*, 4:55, 2010.

[14] B. Finkenstädt, E.A. Heron, M. Komorowski, K. Edwards, S. Tang, C.V. Harper, J.R.E. Davis, M.R.H. White, A.J. Millar, and D.A. Rand. Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics*, 24(24):2901–2907, 2008.

[15] M.K. Oh, L. Rohlin, K.C. Kao, and J.C. Liao. Global expression profiling of acetate-grown *Escherichia coli*. *Journal of Biological Chemistry*, 277(15):13175–13183, 2002.

[16] A. Kolb, S. Busby, H. Buc, S. Garges, and S. Adhya. Transcriptional regulation by cAMP and its receptor protein. *Annual Review of Biochemistry*, 62:749–795, 1993.

[17] G. Gosset, Z. Zhang, S. Nayyar, W.A. Cuevas, and M.H. Saier Jr. Transcriptome analysis of Crp-dependent catabolite control of gene expression in *Escherichia coli*. *Journal of Bacteriology*, 186(11):3516–3524, 2004.

[18] R.M. Gutierrez-Ríos, J.A. Freyre-Gonzalez, O. Resendis, J Collado-Vides, M. Saier Jr, and G. Gosset. Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli*. *BMC Microbiology*, 7:53, 2007.

[19] S. Berthoumieux, H. de Jong, G. Baptist, C. Pinel, C. Ranquet, D. Ropers, and J. Geiselmann. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Molecular Systems Biology*, 9:634, 2013.

[20] T. Kuhlman, Z. Zhang, M.H. Saier Jr., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the USA*, 104(14):6043–8, 2007.

[21] A. Zaslaver, A. Bren, M. Ronen, S. Itzkovitz, I. Kikoin, S. Shavit, W. Liebermeister, M.G. Surette, and U. Alon. A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nature Methods*, 3(8):623–628, 2006.

[22] A Kremling. Comment on mathematical models which describe transcription and calculate the relationship between mrna and protein expression ratio. *Biotechnology and Bioengineering*, 96(4):815–819, 2007.

[23] Nadya Morozova, Andrei Zinovyev, Nora Nonne, Linda-Louise Pritchard, Alexander N Gorban, and Annick Harel-Bellan. Kinetic signatures of microrna modes of action kinetic signatures of microrna modes of action. *Rna New York Ny*, pages 1635–1655, 2012.

[24] J.-C. Leloup and A. Goldbeter. Toward a detailed computational model for the mammalian circadian clock. *Proceedings of the National Academy of Sciences of the USA*, 100(12):7051–7056, 2003.

# Session 8: Evolution

## Conférence invitée

LAURENT DURET

Laboratoire Biométrie et Biologie Evolutive
UMR CNRS 5558, Université Lyon 1
Villeurbanne, France
Laurent.Duret@univ-lyon1.fr
http://lbbe.univ-lyon1.fr/-Duret-Laurent-.html?lang=en

### Meiotic recombination and the evolution of the human genome

Recombination is typically thought as a symmetrical process resulting in large-scale reciprocal genetic exchanges between homologous chromosomes. Recombination events, however, are also accompanied by short-scale, unidirectional exchanges in the neighborhood of the initiating double-strand break: gene conversion. A large body of evidence suggests that gene conversion is GC-biased in many eukaryotes, including mammals and human. AT/GC heterozygotes produce a larger amount of GC- than AT-gametes, thus conferring a population advantage to GC-alleles in high-recombining regions. This apparently unimportant feature of our molecular machinery has strong evolutionary consequences. Structurally, GC-biased gene conversion explains the spatial distribution of GC-content in mammalian genomes - the so-called isochore structure. Functionally, GC-biased gene conversion promotes the "undesired" segregation and fixation of deleterious AT®GC mutations, thus increasing our genomic mutation load. I will review the recent evidence for a GC-biased gene conversion process in mammals, its consequences on genomic landscapes, molecular evolution, and human functional genomics.

# Comparative analysis of phylogenetic profiles for the enzymatic characterization of fungal groups

Cécile PEREIRA[1,2], Jérôme AZÉ[2,3], Alain DENISE[1,2,3], Christine DREVET[1], Christine FROIDEVAUX[2,3], Philippe SILAR[1,4], Olivier LESPINET[1,2]

[1] IGM, UMR 8621 CNRS, Université Paris-Sud, Bât 400, 91405 Orsay Cedex, France
cecile.pereira@igmors.u-psud.fr
[2] LRI, UMR 8623 CNRS, Université Paris-Sud, Bât 650, 91405 Orsay Cedex, France
[3]INRIA AMIB, Saclay, France
4 LIED, FRE, Université Paris Diderot, Sorbonne Paris Cité, Paris, France

**Abstract** *We try to characterize the evolutionary origin of the enzymatic repertoire of different fungal groups. The characteristics for each of the groups studied are determined through the application of* data mining *method on enzyme profiles previously determined by comparative genomics. Through the presentation of results for taxonomic groups Agaricomycetes and Pezizomycota, we show that the application of supervised learning methods is effective in extracting information from phylogenetic* profiles. *We extract specific enzyme activities combinations for each taxonomic groups covered by our analysis. Our approach also enables us to highlight the existence of probable horizontal gene transfers.*

**Keywords** Data Mining, fungi, phylogenetic profiles, evolution, enzymes.

## Étude comparative des profils phylogénétiques dans le but de définir les spécificités enzymatiques de différents groupes de champignons.

**Résumé** *Nous essayons de caractériser l'origine évolutive du répertoire enzymatique de différents groupes de champignons. Les caractéristiques de chacun des groupes étudiés sont déterminées grâce à l'application de méthodes de* fouille de données *sur les* profils enzymatiques *préalablement déterminés par génomique comparée. À travers la présentation des résultats obtenus pour les groupes taxonomiques des Agaricomycetes et des Pezizomycota, nous montrons que l'application de méthodes d'apprentissage supervisé est efficace pour extraire de l'information des* profils phylogénétiques. *Notre approche permet également de mettre en évidence l'existence de probables transferts horizontaux.*

**Mots-clés** Fouille de données, champignons, profils phylogénétiques, évolution, enzymes.

## 1 Introduction

Les champignons possèdent un vaste répertoire enzymatique leur permettant de dégrader et de synthétiser de nombreux composés organiques. Nous nous intéressons à l'origine évolutive de ce répertoire. Plus précisément, notre but est de caractériser les spécificités enzymatiques de différents groupes de champignons. Ceux-ci sont constitués soit à partir de critères taxonomiques, soit à partir de critères ayant trait aux modes de vie.

La comparaison des voies métaboliques de différents organismes a déjà fait l'objet de plusieurs travaux, ainsi ce type d'analyse a déjà permis de reconstruire des arbres phylogénétiques cohérents, preuve de la persistance d'information évolutive dans ces voies [1], la topologie des voies métaboliques a également été étudiée dans le but de proposer de nouvelles phylogénies [2] enfin, de nouvelles cibles thérapeutiques ont pu être définies grâce à la comparaison des voies métaboliques d'organismes pathogènes et non pathogènes [3]. Cependant aucune étude portant sur la totalité du répertoire enzymatique de plus d'une centaine d'espèces n'a encore été menée à ce jour.

Dans ce travail, nous essayons de déterminer la spécificité du répertoire enzymatique des différents groupes de champignons par apprentissage supervisé à partir de profils enzymatiques préalablement établis par détection des gènes homologues entre 165 espèces de champignons [4] (Tableau supplémentaire 1).

Cette approche nous a permis de définir quelles étaient les activités enzymatiques caractéristiques par exemple des *Agaricomycetes* et des *Pezizomycotina*. Elle a également permis de mettre en évidence de probables transferts horizontaux entre plusieurs des espèces ou groupes étudiés.

## 2 Méthodologie

### 2.1 Les données

Nous travaillons à partir de 165 espèces eucaryotes complètement séquencées dont 161 espèces de champignons (Table supplémentaire 1). Les espèces ont été choisies en fonction soit de leur position taxonomique (de manière à échantillonner l'ensemble de la diversité des *Eumycota*), soit de leurs caractéristiques biologiques (de manière à couvrir des modes de vies et d'habitats différents).

La comparaison exhaustive des 1 748 866 protéines constitutives des 165 espèces nous a permis de constituer 139 004 groupes de protéines homologues. Tous les groupes ont été annotés avec le même protocole d'annotation fonctionnelle [4] afin d'éliminer d'éventuels biais liés aux protocoles d'annotation initialement utilisés pour chacun des génomes. Nous obtenons ainsi 12 505 groupes possédant une annotation fonctionnelle de type enzymatique caractérisée par un ou plusieurs *Enzyme Commission* (EC) *numbers* [5]. 1 412 *EC numbers* différents sont utilisés pour définir la fonction de ces 12 505 groupes.

À partir de la distribution des EC numbers nous construisons des profils enzymatiques et des profils phylogénétiques. Le profil enzymatique d'un génome donné est défini par la liste des activités enzymatiques présentes dans ce génome. De la même façon, le profil phylogénétique d'une activité enzymatique donnée est définie par la liste des génomes qui possèdent un gène portant cette activité enzymatique. Ces deux types de profils peuvent être représentés par une matrice (Table 1) à deux dimensions dans laquelle chaque case indique la présence ou l'absence d'une activité enzymatique donnée dans une espèce donnée. Les 1 412 EC *numbers* différents retrouvés parmi les 12 505 groupes d'homologues présentant une activité enzymatique peuvent se répartir en 1 155 profils distincts puisque plusieurs EC *numbers* peuvent présenter un même profil.

| Génome \ EC number | 1.1.1.1 | 1.1.1.108 | 1.1.1.116 | 1.1.1.138 | 1.1.1.14 | 1.1.1.157 | 1.1.1.158,4.3.1.2 |
|---|---|---|---|---|---|---|---|
| ***Tuber melanosporum*** | | | # | | | | |
| ***Arthrobotrys oligospora*** | | | # | | | | |
| ***Aspergillus nidulans*** | | | # | | | | |
| ***Talaromyces stipitatus*** | | | # | | | | |
| ***Penicillium marneffei*** | | | # | | | | |
| ***Penicillium chrysogenum*** | * | * | # * | * | * | * | * |
| ***Neosartorya fischeri*** | | | # | | | | |
| ***Aspergillus aculeatus*** | | | # | | | | |
| ***Aspergillus flavus*** | | | # | | | | |

**Table 1 :** Exemple de profils enzymatiques et de profils phylogénétiques. Les génomes sont disposés en ligne et les EC *numbers* en colonne (listes non exhaustives). Lorsque une activité enzymatique donnée est présente chez une espèce donnée la case du tableau est colorée en gris. Le profil enzymatique de *Penicillium chrysogenum* est symbolisé par des '*'. Le profil phylogénétique de l'EC 1.1.1.116 est symbolisé par des '#'.

### 2.2 Apprentissage supervisé

Nous utilisons des méthodes d'apprentissage supervisé afin d'extraire des informations sur la spécificité du répertoire enzymatique de différents groupes de champignons à partir des profils phylogénétiques.

Il existe un large panel de méthodes d'apprentissage de natures différentes (modèle bayésien, arbres de décision, règles de classification, k-plus proches voisins, *etc*.). Nous avons choisi d'utiliser des approches fournissant des modèles de classification interprétables et à fort pouvoir explicatif tels que des arbres de décision ou des règles de classification. Les arbres de décision sont des approches *top-down*, c'est-à-dire qu'à chaque étape, le triplet (attribut, test, valeur) qui optimise le critère est retenu et deux partitions disjointes sont créées. Les règles utilisent quant à elles une approche *bottom-up*, c'est-à-dire effectuent une généralisation à partir d'un exemple de manière à couvrir le plus possible d'exemples positifs, sans couvrir d'exemples négatifs. Afin d'exploiter les profils phylogénétiques nous combinons l'utilisation des trois algorithmes d'apprentissage supervisés C4.5 [4] (arbre de décision), PART [5] (règle de décision) et RIPPER [6] (règles construites directement) avec un système de vote majoritaire. Le critère d'évaluation choisi pour ces 3 algorithmes est le gain d'information. L'implémentation de ces algorithmes est fournie par la boîte à outils WEKA [7].

Les méthodes d'apprentissage supervisé appliquées sur les champignons décrits par les profils enzymatiques renvoient comme résultat des classifieurs permettant de mettre en évidence les combinaisons d'enzymes spécifiques d'un groupe taxonomique donné. En d'autres termes, elles permettent de mettre en évidence les « synapomorphies enzymatiques » d'un groupe d'espèces. Dans ce but, nous définissons les exemples positifs comme étant l'ensemble des génomes appartenant à un groupe donné, et les exemples négatifs comme étant l'ensemble des génomes n'appartenant pas à ce groupe (apprentissage de deux classes).

Afin d'estimer la qualité des classifieurs obtenus par chacune des trois méthodes de classification choisies, nous avons utilisé la méthode « *Leave One Out* ». Il est à noter que nous avons prédit les groupes d'orthologues à partir de la totalité des génomes. Ainsi, *stricto sensu,* il existe un lien entre les données d'apprentissage et de test, ce qui peut être une source de biais dans l'évaluation. Cependant, ce n'est pas sur les groupes d'orthologues que ce fait l'évaluation, mais sur les profils phylogénétiques. Ces profils ont été obtenus par l'annotation des groupes d'orthologues. Ces groupes sont annotés indépendamment et plusieurs groupes peuvent porter la même annotation. Ainsi l'utilisation des profils permet une diminution de ce possible biais.

A)

| **Réel (R)\ Prédit (P)** | **Positif** | **Négatif** |
|---|---|---|
| **Positif** | Vrais positifs (VP) | Faux négatifs (FN) |
| **Négatif** | Faux positifs (FP) | Vrais négatifs (VN) |

B)

$$\text{Précision} = \frac{VP}{VP + FP}$$

$$\text{Sensibilité} = \frac{VP}{VP + FN}$$

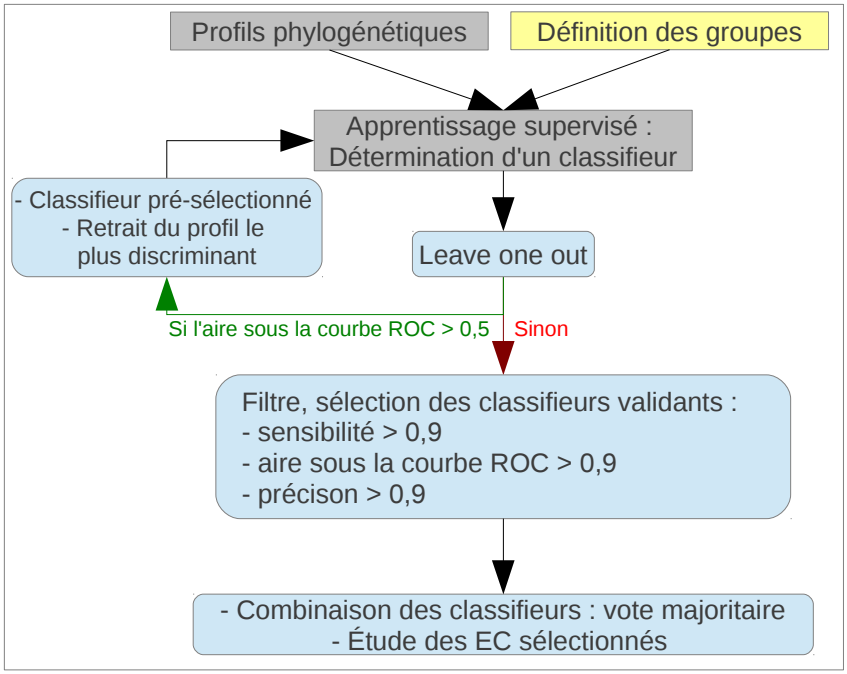
**Figure 1 :** (A)Matrice de confusion et (B) Critères d'évaluation associés.

La matrice de confusion (Figure 1A) est obtenue en comparant le groupe taxonomique associé à chaque champignon de l'ensemble de test (méthode *Leave One Out*) avec le groupe taxonomique de référence de cette espèce. Elle se construit en mettant respectivement sur les lignes et sur les colonnes les groupes taxonomiques de référence et la classification faite par le classifieur.

La majorité des groupes étudiés ont un nombre d'exemples positifs très inférieur au nombre d'exemples négatifs, c'est pourquoi nous avons fait le choix de critères d'évaluation décrivant majoritairement la qualité de prédiction des exemples positifs. Nous évaluons donc la qualité des classifieurs en fonction de trois critères différents (Figure 1B) : la sensibilité, la précision et l'aire sous la courbe ROC (*Receiver Operating Characteristic*) . Ces critères ont des valeurs comprises entre 0 et 1. Les classifieurs d'un groupe taxonomique sont conservés s'ils ont, pour ces trois critères, une valeur supérieur à 0,9.

Nous cherchons à caractériser un groupe d'organismes en fonction des activités enzymatiques. Or il est possible que plusieurs combinaisons d'activités enzymatiques différentes soient pertinentes pour caractériser ce groupe. De ce fait, afin d'obtenir un large panel de classifieurs pertinents nous appliquons de façon itérée la recherche de classifieurs sur les données. La recherche des classifieurs s'effectue en deux étapes. La première consiste à collecter l'ensemble des classifieurs meilleurs que des classifieurs aléatoires (aire sous la courbe ROC supérieure à 0,5). Le seconde étape à pour objectif de filtrer les classifieurs en fonction dess

seuils fixés pour la sensibilité, la précision et l'aire sous la courbe ROC. Les profils phylogénétiques sélectionnés en premier par l'algorithme d'apprentissage sont retirés du jeu de données et un nouveau classifieur est appris (Figure 2).




Pour un groupe taxonomique donné, les classifieurs obtenus avec les trois approches prédisent chacun l'appartenance du champignon à un groupe (le groupe taxonomique en cours de caractérisation ou « autre », c'est-à-dire tout sauf ce groupe). L'ensemble des activités enzymatiques sélectionnées par les classifieurs forment les activités enzymatiques caractéristiques de ce groupe. La prédiction du groupe taxonomique d'un nouveau champignon sera faite en conservant la prédiction majoritaire de l'ensemble des classifieurs finaux.

**Figure 2 :** Pipeline d'apprentissage supervisé pour un algorithme d'apprentissage donné. Les différentes étapes sont décrites sur les rectangles aux angles arrondis à fond bleu, les données que nous générons sont représentées sur fond gris et les données de la littérature sur fond jaune.

## 3 Résultats

D'une manière générale les résultats de l'application de cette méthodologie ont donné des classifieurs faisant intervenir un faible nombre d'activités enzymatiques avec de forts pourcentages de bonne classification. L'obtention de ces classifieurs s'effectue en un temps raisonnable, l'apprentissage d'un classifieur pour un groupe donné (une itération de la boucle de la figure 2) prenant moins d'une minute.

Nous traitons ici en tant qu'exemple la caractérisation des *Agaricomycetes* (phylum des *Basidiomycota*) à partir de l'ensemble des espèces présentes dans nos données ainsi qu'une partie de la caractérisation des *Pezizomycotina (*phylum *Ascomycota)* parmi les champignons.

### 3.1 Caractérisation des Agaricomycetes

L'application de notre méthode (Figure 2) à la détermination des *Agaricomycetes* a permis de construire les 13 classifieurs de la table supplémentaire 2. Les profils phylogénétiques sélectionnés par les classifieurs sont présentés dans la table 2. Les activités enzymatiques sélectionnées par les classifieurs comme caractéristiques des *Agaricomycetes* sont impliquées dans différentes processus biologiques que nous décrivons ci-dessous.

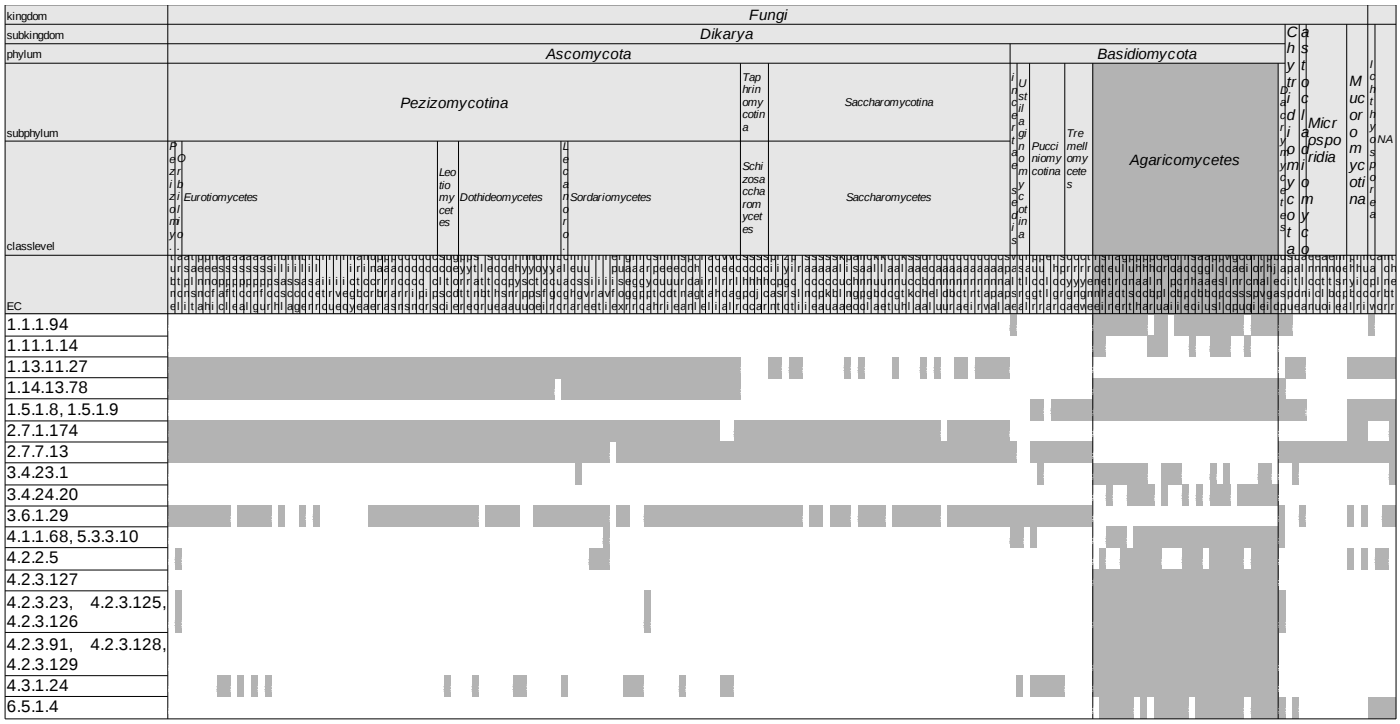



**Table 2** : Profils phylogénétiques des activités enzymatiques présentes dans les classifieurs caractérisant la classe *Agaricomycetes*. Les activités enzymatiques sont disposées en ligne et les génomes sont disposés en colonne. Les cases sur fond gris signifient que l'activité enzymatique est présente dans le génome, sur fond blanc qu'elle est absente. Le groupe taxonomique des *Agaricomycetes* est sur fond gris foncé.

### 3.1.1 Création de composés biologiquement actifs

Les EC :4.2.3.91 (cuberol synthase), EC :4.2.3.127 (beta-copaenz synthase), EC :4.2.3.128 (beta-cubebene synthase) et 4.2.3.129 ((+)-sativene synthase) ont été retrouvés caractéristiques des *Agaricomycetes* par l'ensemble des méthodes. Elles correspondent aux activités enzymatiques de la Sesquiterpene synthase COP4 connue pour être impliquée dans la catalyse de la cyclisation du farnesyl diphosphate en plusieurs produits, incluant la germacrene D, la beta-copaene, la bete-cubebene, la (+)-sativene et le cuberol. Elles catalysent la formation de terpenoids intermédiaires à la création de composés biologiquements actifs tels que les antibiotiques et les toxines [13]. De plus, l'activité enzymatique EC :1.14.13.78 (ent-kaurene oxidase) retrouvée caractéristique des *Agaricomycetes*, a elle aussi un rôle dans la synthèse des terpenoides (gibberellins). Les *Agaricomycetes* se caractérisent donc en partie par leur capacité à produire de tels types de composés.

Trois autres activités enzymatiques correspondant aux activités enzymatiques de la Sesquiterpene synthases COP3 sont retrouvées, il s'agit des EC : 4.2.3.126 (Alpha-muurolene synthase), EC : 4.2.3.23 (Gamma-muurolene synthase) et EC : 4.2.3.125 (Germacrene-A synthase). Les produits des réactions catalysées par ces enzymes sont des composés (ou des intermédiaires de composés) biologiquement actifs. Le germacrene est par exemple un agent antimicrobien.

### 3.1.2 Dégradation de la biomasse

Les activités enzymatiques EC :1.11.1.14 (lignine peroxidase) ,EC :3.4.24.20 (Peptidyl-Lys metalloendopeptidase), EC :4.2.2.5 (chondroitine AC lyase) et EC:3.4.23.1 (pepsin A) sont retrouvées comme caractéristiques des *Agaricomycetes*. La peptidyl-Lys metalloendopeptidase est une protéase sécrétée. La pepsin A est une endopeptidase. La lignin peroxidase est impliquée dans la dégradation de la lignine [13]. La chondroitine AC lyase permet la dépolymérisation de la chondroitine sulfate (constituant du

cartilage) et du dermatan sulfate (constituant de la peau), ainsi cette activité enzymatique pourrait permettre la dégradation de la biomasse d'origine animale.

Le groupe des *Agaricomycetes* contient à la fois les pourritures brunes et les pourritures blanches. Elles sont toutes deux connues pour leur capacité de dégradation de la biomasse. Ainsi trouver ces activités enzymatiques caractéristique de ce groupe est cohérent avec les connaissances actuelles sur ces organismes.

### 3.1.3 Paroi et membrane cellulaire

Une activité enzymatique spécifiquement présente au niveau de la paroi et de la membrane cellulaire a été sélectionnée. Il s'agit de activité enzymatique EC :2.7.7.13 (mannose-1-phosphate guanylyltransferase).

L'absence de la mannose-1-phosphate guanylyltransferase est caractéristique des *Agaricomycetes*. Cette activité enzymatique permet la formation de GDP-mannose, lui même impliqué dans la formation de la paroi. Ainsi la paroi des Agaricomycetes semble être particulière ou sa synthèse implique d'autres activités enzymatiques (sous-voies différentes) induisant la production de GDP-mannose.

### 3.1.4 Metabolisme des nucléotides

Les *Agaricomycetes* se caractérisent également comme ne possédant pas l'activité enzymatique EC : 3.6.1.29 (bis(5'-adenosyl)-triphosphatase). Cette activité enzymatique est en lien avec le métabolisme des nucléotides. Ainsi ce groupe d'organisme semble posséder un métabolisme des nucléotides particulier différent de celui des autres *Basidiomycetes*.

### 3.1.5 Stockage de glycerolypides

L'activité enzymatique EC :1.1.1.94 (glycerol-3-phosphate dehydrogenase [NAD(P)+]) est impliquée dans le stockage des triglycérides. Les *Agaricomycetes* ne sont pas spécialement connus pour avoir cette capacité. Ainsi il s'agit donc soit d'une erreur d'annotation soit de la découverte d'une nouvelle capacité caractéristique de se groupe de champignons qu'il serait intéressant de tester.

### 3.1.6 Métabolisme des vitamines

L'absence de la mannose-1-phosphate guanylyltransferase (EC: 2.7.7.13) chez les *Agaricomycetes* pourrait également avoir des conséquences dans les possibilités de biosynthèse de la vitamine C.

### 3.1.7 Métabolisme de acides aminés

Plusieurs activités enzymatiques sélectionnées sont impliquées dans le métabolisme de la tyrosine et de la phenylalanine. Les activités enzymatiques EC :4.1.1.68 et EC :5.3.3.10 agissent de manière consécutive dans la métabolisme de la tyrosine. Elles appartiennent à une sous-voie pouvant avoir comme précurseur la phenylalanine ou la tyrosine. L'EC :4.3.1.24 (phenylalanine ammonia-lyase) permet la formation du trans-Cinnamate, substrat initial d'une partie des voies de dégradation de la phenylalanine. Enfin, l'activité enzymatique EC :1.13.11.27 permet la transformation du phenylpyruvate en 2-hydroxy-phenylacetate, précurseur de la voie de dégradation des styrenes.

Les activités enzymatiques EC:1.5.1.8 (saccharopine dehydrogenase (NADP+, L-lysine-forming)) et EC:1.5.1.9 (saccharopine dehydrogenase (NAD+, L-glutamate-forming)) ont un rôle consécutif dans le métabolisme de la lysine. Elles permettent l'entrée possible de la lysine vers les le cycle du citrate (dégradation) et vers la voie de production de penicilline et cepholosporine.

Les mécanismes de modification et de dégradation de certains acides aminés semble donc caractéristique des *Agaricomycetes.*

### 3.1.8 Regulation de la structure de la membrane nucléaire

La diacylglycerol kinase (CTP dependant) est connue chez *Saccharomyces cerevisiae* pour réguler la

synthèse de phospholipides et la croissance de la membrane nucléaire [15]. Son absence est caractéristique du groupe des *Agaricomycetes*. Ainsi, ces champignons utilisent donc certainement d'autres mécanismes pour effectuer cette régulation.

### 3.2 Caractérisation des *Pezizomycotina*

La caractérisation des *Pezizomycotina* sur les données constituées de l'ensemble des champignons permet la sélection de 381 classifieurs. Parmi ceux-ci nous ne présenterons ici que le premier classifieur obtenu par les 3 algorithmes. Ce classifieur est le suivant : si l'activité enzymatique 3.1.6.6 est présente alors le champignon appartient au groupe des *Pezizomycotina* sinon il appartient à un autre groupe. L'EC 3.1.6.6 (choline-sulfatase) intervient dans la voie de dégradation de la choline-o-sulfate, permettant une source de soufre endogène mobilisable pendant la croissance [16]. Les *Pezizomycotina* se caractérisent donc en partie par leur capacité à croître sur un milieu pauvre en soufre.

La choline-sulfatase est présente chez l'ensemble des *Pezizomycotina* ainsi que chez *Ustilago maydis* (un *Basidiomycota*). Nous cherchons à comprendre d'où provient cette activité enzymatique chez U. *maydis.* Au total, 85 protéines annotées EC:3.1.6.6 appartiennent au même groupe d'orthologue que la séquence d'*U. Maydis*. Sur l'arbre phylogénétique obtenu à partir de ces 85 séquences (Figure 3) la séquence correspondant à la protéine de *U. maydis* n'est pas sur une branche externe mais est nichée proche d'un groupe de séquences de *Pezizomycotina*. Cela conduit à penser que la présence de l'activité enzymatique EC:3.1.6.6 est probablement due à un transfert horizontal entre une souche de *Pezizomycotina* et *U. maydis.*

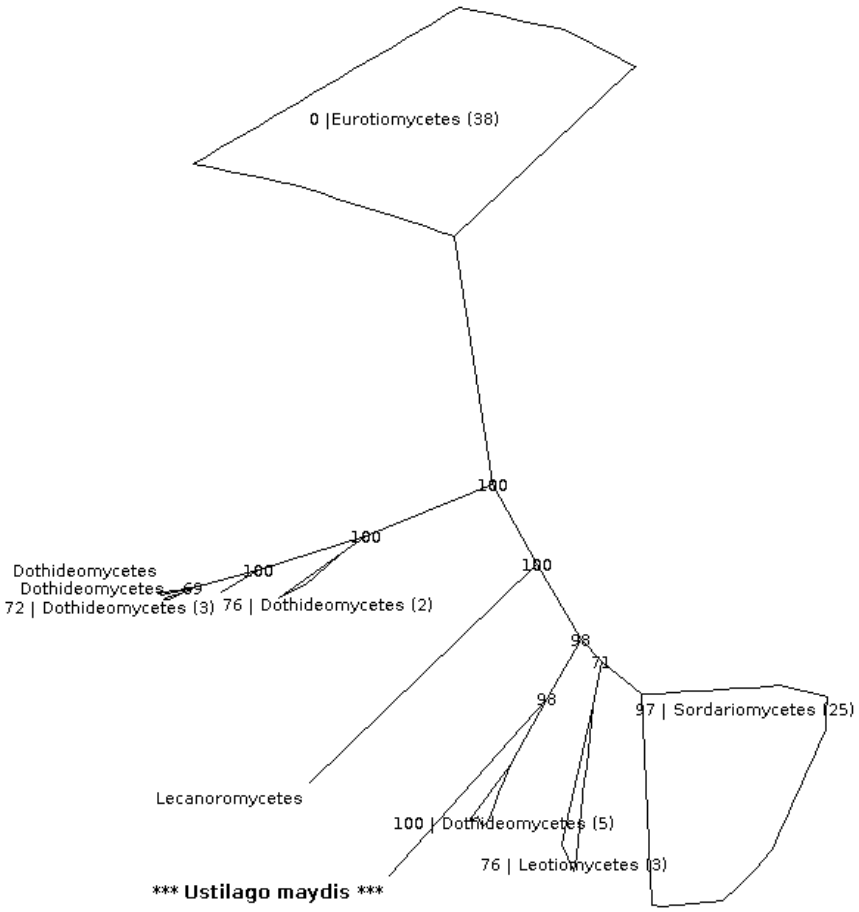



**Figure 3** : Arbre phylogénétique construit avec le groupe d'orthologue de la sequence d'*U. maydis* annoté 3.1.6.6. L'identifiant de la sequence de *U. Maydis* est représentée en gras et encadré par des étoiles ('***'). Pour les autres séquences seul la classe taxonomique est indiquée. Les séquences ont été alignées avec

MUSCLE [8] et les parties les moins bien alignées ont été retirées à l'aide du logiciel Gblocks [9]. L'arbre phylogénétique a été construit à l'aide du programme PhyML [10]. Les nombres indiqués sur certain nœuds correspondent à la valeur de bootstrap.

## 4 Conclusion

L'application de méthodes d'apprentissage supervisé sur les groupes taxonomiques à partir de profils enzymatiques nous permet de caractériser un groupe d'organismes en fonction de ses activités enzymatiques. À travers la caractérisation des *Agaricomycetes* et des *Pezizomycotina*, nous mettons en évidence la validité de l'application de méthodes d'apprentissage supervisé dans le but d'extraire de l'information des profils phylogénétiques. Cette méthode permet également de poser de nouvelles hypothèses sur l'évolution du répertoire enzymatique de ces organismes, notamment en détectant des transferts horizontaux.

Par la suite, cette méthodologie sera appliquée à la caractérisation d'autres groupes partageant une caractéristique commune dans le but de comprendre les mécanismes induisant cette caractéristique. Nous l'appliquerons par exemple à la caractérisation des champignons ayant de fortes capacités de dégradation de la biomasse. L'étude de la biomasse se fera sur les profils phylogénétiques des activités enzymatiques ainsi que sur l'étude des groupes d'orthologues afin de trouver de nouvelles protéines impliquées dans ce processus. Nous pourrons ainsi prédire les capacités de dégradation de la biomasse de nouvelles espèces de champignons.

## Remerciments

Ce travail a bénéficié d'un financement par le PEPS Bio-Maths-Info (BMI) CNRS-INSERM-INRIA .

## References

[1] Y. Zhang, S. Li, G. Skogerbo, Z. Zhang, X. Zhu, Z. Zhang, S. Sun, H. Lu, B. shi and R. Cher, Phylogenetic properties of metabolic pathway topologies as revealed by global analysis. *BMC Bioinformatics*, doi:10.1186/1471-2105-7-252, 2006.

[2] A. Mano, T. Tuller, O. Béjà and R. Y. Pinter, Comparative classification of species and the study of pathway evolution based on the alignment of métabolic pathways. *BMC Bioinformatics*, doi:10.1186/1471-2105-11-S1-S38, 2010

[3] D. Perumal, C. S. Lim and M. K. Sakharkar, A comparative study of metabolic network topology between a pathogenic and a non-pathogenic bacterium for potential drug target identificaion. *Translat Bioinforma*, 2009: 100-104, 2009

[4] S. Grossetête, B. Labedan and O. Lespinet, FUNGIpath, a tool to assess fungal metabolic pathways predicted by orthology. *BMC Genomics*, doi:10,1186/1471-2164-11-81, 2010.

[5] IUPAC-IUBMB. IUBMB Joint Commission on Biochemical Nomenclature (JCBN) and Nomenclature Committee of IUBMB, Newsletter 1999. *Eur. J Biochem.* 1999;264:607-609

[6] J.R. Quinlan, C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers*, 1993.

[7] E. Frank, I. H. Witten, Generating Accurate Rule Sets Without Global Optimization. *Fifteenth International Conference on Machine Learning*, 144-151, 1998.

[8] W. W. Cohen, Fast Effective Rule Induction. *Twelfth International Conference on Machine Learning*, 115-123, 1995.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update; *SIGKDD Exploration*, Volume 11, Issue 1, 2009

[10] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Res.* **32**(5):1792-1797, 2004

[11] J. Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17, 540-552. 2000

[12] S. Guindon, J.F. Dufayard, V. Lefort, M. Anisimova, W. Anisimova, O. Gascuel, New algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3):307-21, 2010

[13] F. Lopes-Gallego, S. A. Agger, D. A. Pella, M.D. Distefano, C. Schmidt-Dannert, Sesquiterpene synthase Cop4 and Cop6 from *Coprinus cinereus*: Catalytic promiscuity and cyclization of farnesyl pyrophosphate geometrical isomers. *Chembiochem*, 11(8):1093-1106, 2010

[14] G. Daniel, J. Volc, L. Filonova, O. Plíhal, E. Kubátová, P. Halada. Characteristics of Gloeophyllum trabeum alcohol oxidase, an extracellular source of H2O2 in brown rot decay. *Appl Environ Microbiol 73(19):6241-53*, 2007.

[15] G. Han, L. O'Hara, G. M. Carman and S. Siniossoglou. An unconventional diacylglycerol kinase that regulates phospholipid synthesis and nulcear membrane growth. *J Biol Chem. 283(29):20433-20442*, 2008.

[16] R. A. Gravel. Choline-O-sulphate utilization in *Aspergillus nidulans. Genetical Research, 28(3):261-76, 1976*

## Annexes

| Groupe taxonomique | | | | | | nombre de génomes |
|---|---|---|---|---|---|---|
| Eumycota | Dikarya | Ascomycota | Pezizomycotina | Dothideomycetes | Capnodiales | 5 |
| | | | | | Hysteriales | 2 |
| | | | | | Pleosporales | 8 |
| | | | | Eurotiomycetes | Eurotiales | 13 |
| | | | | | Onygenales | 24 |
| | | | | Lecanoromycetes | Lecanorales | 1 |
| | | | | Leotiomycetes | Helotiales | 3 |
| | | | | Orbiliomycetes | Orbiliales | 1 |
| | | | | Pezizomycetes | Pezizales | 1 |
| | | | | Sordariomycetes | Diaporthales | 1 |
| | | | | | Glomerellales | 5 |
| | | | | | Hypocreales | 8 |
| | | | | | Magnaporthales | 3 |
| | | | | | Sordariales | 8 |
| | | | Saccharomycotina | Saccharomycetes | Saccharomycetales | 35 |
| | | | Taphrinomycotina | Schizosaccharomycetes | | 4 |
| | | Basidiomycota | Basidiomycota incertae sedis | Wallemiomycetes | Wallemiales | 1 |
| | | | Agaricomycetes | | Agaricales | 7 |
| | | | | | Auriculariales | 1 |
| | | | | | Boletales | 2 |
| | | | | | Corticiales | 4 |
| | | | | | Gloeophyllales | 1 |
| | | | | | Hymenochaetales | 1 |
| | | | | | Polyporales | 9 |
| | | | | | Russulales | 2 |
| | | | Dacrymycetes | | Dacrymycetales | 1 |
| | | | Tremellomycetes | | Tremellales | 4 |
| | | | Pucciniomycotina | | Pucciniales | 3 |
| | | | Pucciniomycotina | | Sporidiobolales | 2 |
| | | | Ustilaginomycotina | Exobasidiomycetes | Malasseziales | 1 |
| | | | | Ustilaginomycetes | Ustilaginales | 1 |
| | Blastocladiomycota | | Blastocladiomycetes | | Blastocladiales | 1 |
| | Chytridiomycota | | Chytridiomycetes | | Spizellomycetales | 1 |
| | | | | | Rhizophydiales | 1 |
| | Microsporidia | | | | | 6 |
| | Mucoromycotina | | | | Mucorales | 3 |
| Apusozoa | | | | | | 1 |
| Filasterea | | | | | | 1 |
| Choanoflagellida | | | | | | 2 |

**Table supplémentaire 1 :** Distribution taxonomique des génomes. Les groupes taxonomiques de champignons sont sur fond gris, les autres sur fond blanc.

| Algorithmes | Données | Classifieurs |
|---|---|---|
| **C4.5** | Tous les profils | Si les EC : 4.2.3.127 et 4.2.3.91 et 4.2.3.128 et 4.2.3.129 sont présents alors *Agaricomycetes* (151/0) (un seul profil)<br>Sinon autre (27/0) |
| | Tous les profils sauf celui de EC : 4.2.3.127 | Si les EC : 4.2.3.126 et 4.2.3.23 et 4.2.3.125 sont absents alors autre (148.0/0) (un seul profil)<br>Sinon si l'EC : 4.3.1.24 est absent autre (3.0)<br>Sinon *Agaricomycetes* (27.0) |
| | Tous les profils sauf ceux de EC : 4.2.3.127 et EC :4.2.3.126 | Si les EC : 2.7.7.13 et EC : 1.5.1.9 et EC : 1.5.1.8 sont absents alors autre (3.0/0)<br>Sinon si l'EC : 2.7.7.13 est absent et que les EC :1.5.1.9 et 1.5.1.8 sont présents alors (27.0)<br>Sinon autre (148.0) |
| | 7 profils phylogénétiques retirés | Si les EC : 4.1.1.68, EC : 5.3.3.10 et EC : 1.11.1.14 sont absents alors autre (147.0/1.0)<br>Sinon, si les EC : 4.1.1.68, EC : 5.3.3.10 sont absents et que l'EC : 1.11.1.14 est présent alors *Agaricomycetes* (2.0/0)<br>Sinon si les EC : 4.1.1.68 et EC : 5.3.3.10 sont présents et que l'EC : 3.6.1.29 est absent alors *Agaricomycetes* (24.0/0)<br>Sinon autre (5.0/0) |
| **PART** | Tous les profils | Si EC : les EC 4.2.3.127 et 4.2.3.91 et 4.2.3.128 et 4.2.3.129 sont présents alors *Agaricomycetes*<br>Sinon autre |
| | Tous les profils sauf celui de EC : 4.2.3.127 | Si les EC : 4.2.3.126 et 4.2.3.23 et 4.2.3.125 sont absents alors autre (148.0/0) (un seul profil)<br>Sinon si l'EC :4.3.1.24 est absent autre (3.0)<br>Sinon *Agaricomycetes* (27.0) |
| | Tous les profils sauf ceux de EC : 4.2.3.127 et EC :4.2.3.126 | Si l'EC :2.7.7.13 est présent alors autre (148.0/0)<br>Sinon, si l'EC :1.5.1.9 et l'EC : 1.5.1.8 sont présents alors *Agaricomycetes* (27.0/0) (même profil)<br>Sinon autre (3.0/0) |
| | 5 profils phylogénétiques retirés | Si l'EC :1.1.1.94 et l'EC : 1.11.1.14 sont absents alors autre (148.0/1.0)<br>Sinon, si l'EC :4.3.1.24 est présent alors *Agaricomycetes* (26.0/0)<br>Sinon autre (4.0/0) |
| | 7 profils phylogénétiques retirés | Si l'EC :4.1.1.68, l'EC : 5.3.3.10 et l'EC : 1.11.1.14 sont absent alors autre (147.0/1.0)<br>Sinon, si l'EC : 3.6.1.29 est absent alors *Agaricomycetes* (26.0/0)<br>Sinon autre (5.0/0) |
| **RIPPER** | Tous les profils | Si EC : les EC 4.2.3.127 et 4.2.3.91 et 4.2.3.128 et 4.2.3.129 sont présents alors *Agaricomycetes*<br>Sinon autre |
| | 15 profils phylogénétiques retirés | Si les EC :6.5.1.4 et 4.3.1.24 sont présents alors *Agaricomycetes* (22.0/0.0)<br>Sinon si l'EC : 3.4.24.20 est présent alors *Agaricomycetes* (3.0/0.0)<br>Sinon si l'EC : 1.11.1.14 est présent alors Agaricomycetes (2.0/0.0)<br>Sinon autre (151.0/0.0) |
| | 41 profils phylogénétiques retirés | Si l'EC : 2.7.1.174 est absent et que l'EC : 1.14.13.78 est présent alors *Agaricomycetes* (30.0/3.0)<br>Sinon autre (148.0/0.0) |
| | 45 profils phylogénétiques retirés | Si l'EC :4.2.2.5 est présent et que l'EC : 1.13.11.27 est absent alors *Agaricomycetes* (19.0/0.0)<br>Sinon, si l'EC : 3.4.23.1 est présent alors *Agaricomycetes* (10.0/4.0)<br>Sinon autre (149.0/2.0) |

**Table supplémentaire 2 :** Classifieurs caractérisant les Agaricomycetes. Les nombres entre parenthèses indiquent dans le cas où une règle génère des erreurs sur le jeu de données complet combien de génomes sont couverts par la règle puis le nombre d'erreurs.

# Searching for virus phylotypes

François CHEVENET[1,2,3], Matthieu JUNG[1,2,4], Martine PEETERS[4], Tulio de OLIVEIRA[5] and Olivier GASCUEL[1,2]

[1] LIRMM, UMR5506 CNRS – Université Montpellier 2, Montpellier, France
{olivier.gascuel, matthieu.jung}@lirmm.fr

[2] Institut de Biologie Computationnelle (IBC), 95 rue de la Galera, Montpellier, France

[3] MIVEGEC, CNRS 5290, IRD 224, Université Montpellier 1 et 2, Montpellier, France

[4] TransVIHMI, UMI233, IRD – Université Montpellier 1, Montpellier, France
{francois.chevenet, martine.peeters}@ird.fr

[5] Africa Centre for Health and Population Studies, University of KwaZulu-Natal, Durban, South Africa
tdeoliveira@africacentre.ac.za

**Abstract:** *Large phylogenies are being built today to study virus evolution, trace the origin of epidemics, establish the mode of transmission and survey the appearance of drug resistance. However, no tool is available to quickly inspect these phylogenies and combine them with extrinsic traits (e.g., geographic location, risk group, presence of a given resistance mutation), seeking to extract strain groups of specific interest or requiring surveillance. We propose a new method for obtaining such groups, which we call phylotypes, from a phylogeny having taxa (strains) annotated with extrinsic traits. Phylotypes are subsets of taxa with close phylogenetic relationships and common trait values. The method combines ancestral trait reconstruction using parsimony, with combinatorial and numerical criteria measuring tree shape characteristics and the diversity and separation of the potential phylotypes. A shuffling procedure is used to assess the statistical significance of phylotypes. All algorithms have linear time complexity. This results in low computing times, typically a few minutes for the larger data sets with a number of shuffling steps. We analyze a large HIV-1 data set containing >3000 strains of HIV-1 subtype C collected worldwide, where the method shows its ability to recover known clusters and transmission routes, and to detect new ones. This method and companion tools are implemented in an interactive Web interface (www.phylotype.org), which provides a wide choice of graphical views and output formats, and allows for exploratory analyses of large data sets. The detailed description of the method, tools, Web interface and application has been published recently in Bioinformatics* [1].

**Keywords:** *Phylogenetics, phylogeography, parsimony, ancestral reconstruction, HIV.*

## 1 Introduction

Phylogenetic tools are commonly used to study virus evolution, trace the origin of epidemics, establish the mode of transmission, survey the apparition of drug resistances, or determine virus origin in different body compartments. The process involves the construction of phylogenetic trees, their visualization, and their interpretation. Ancestral character reconstruction methods aid the interpretation, as extrinsic traits and their evolution can be mapped on the tree. Parsimony has been one of the very first approaches to reconstruct ancestral characters (*e.g.* MacClade [2]). Despite the number of methods available for the inference of ancestral traits, there is little development for the interpretation of trait-annotated phylogenies. Most of the programs display the reconstructed ancestral states but do not allow for tests on ancestry and taxon clustering. For example, MacClade reconstructs ancestral characters and maps them in the phylogeny, but the resulting annotated tree requires to be interpreted visually.

There is a need for a fast, easy-to-use exploratory tool that can use phylogenies constructed with any of the most popular methods, while providing fast inference of ancestral traits and enabling hypothesis testing and visual data interpretation of evolutionary scenarios. Here, we present the PhyloType method that uses and formally defines the concept of “viral phylotype”.

## 2 Method and implementation

A phylotype is a subset of studied taxa that share a common history. This history is 2-fold. The first component is a rooted phylogeny *T* of all taxa studied, which is one of the inputs of the method. Clades (rooted subtrees) are the standard way to define subsets of taxa with common history from a rooted phylogeny. Here, a phylotype must be included in a clade of *T*, and the root of this clade must be the most recent common ancestor (MRCA) of the members of the phylotype. This MRCA is also called the root of the phylotype. In some cases (*e.g.*, when studying the geographical origin of an epidemic), the common clade will be the entire phylogeny *T*. The common clade and MRCA property define the phylogenetic component of the common history of phylotypes.

The second component of this common history is induced by a set of traits or annotations attached to each of the taxa. For example, these annotations may describe the country of origin, the presence of a given resistance mutation, the risk group, the mode of transmission and so forth. These annotations are provided by the user and are the second input of the method. Beyond phylogeny, the members of a given phylotype share the same evolutionary history regarding the annotations being analyzed. The phylotype root must have a unique annotation, say *A*, which lasted until extant phylotype members that are annotated by *A*. Formally, within a clade with root *r*, every taxon *x* sharing the same annotation *A* as *r*, and for which *A* is conserved along the path from *r* to *x*, belongs to the phylotype defined by *r* and *A* (assuming the MRCA condition is fulfilled). As ancestral annotations are unknown, some inference method must be used. The PhyloType software currently uses parsimony, but other approaches are possible.

Phylotypes may be simple clades, but they may also form hierarchical chains showing a succession of founder events. The approach differs from that commonly used for bacteria, where clades with low genetic diversity are aggregated. Here, a phylotype is associated with a unique annotation, for example, a country. The method to extract phylotypes can be summarized as: (1) inference of ancestral annotations using parsimony; (2) this ancestral reconstruction induces a set of potential phylotypes defined by the MRCA and path-conservation conditions (see above); (3) combinatorial, numerical and statistical criteria are used to select a set of relevant phylotypes from among all potential phylotypes.

All criteria are defined recursively and globally computed for all potential phylotypes in O(*n*) time. For instance, the *Size* criterion simply counts the number of taxa (members) in the potential phylotypes. The *Different* criterion measures the number of exceptions in the potential phylotypes. Let *P* be a potential phylotype defined by clade *C*; *Different* counts the number of sub-clades of *C* with annotations differing from *P*'s annotation. When such a sub-clade is found, it is counted as one, regardless of the number of covered taxa. The *Persistence* criterion measures the extent to which the root annotation *A* of the phylotype is conserved in its descendants. It is equal to the minimum depth where *A* is conserved, among all lineages starting from phylotype root. In total, PhyloType offers twelve combinatorial and numerical criteria.

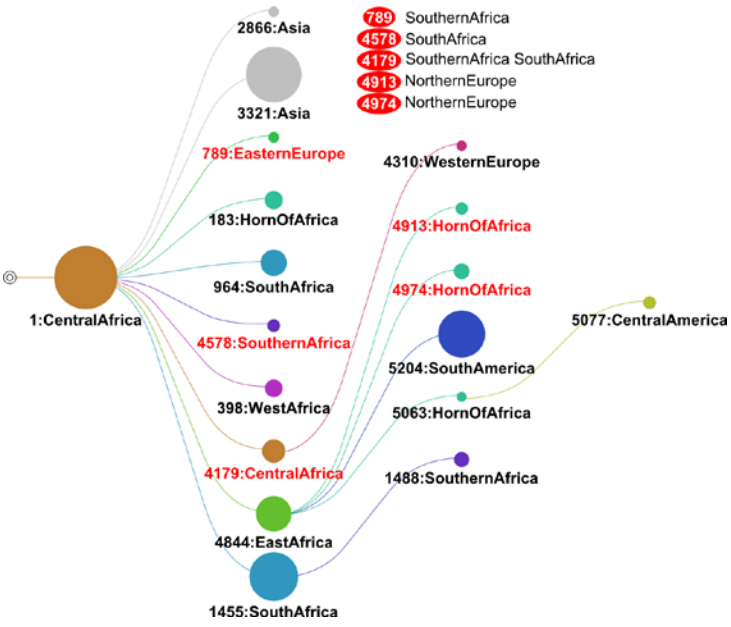
Once all potential phylotypes have been evaluated using the criteria and thresholds defined by the user, we perform a top-down tree traversal to select the most general phylotypes satisfying all criteria. If an *A* phylotype satisfying all criteria is included in another *A* phylotype satisfying all criteria, then only the most general one will be selected, unless the path between the two contains one (or more) node(s) not annotated with *A*; in that case, both *A* phylotypes are selected. The corresponding Search algorithm runs in O(*n*) time.

Once a set of phylotypes has been selected from the data, an essential question is whether or not these phylotypes have some statistical significance and clearly depart from a random selection of taxon subsets within the input phylogeny. For this purpose, we use a shuffling procedure, a common statistical tool that is used for similar purposes in phylogenetic software packages (*e.g.* MacClade) and to study the phylogeography of virus epidemics (*e.g.* [3]). Here, we randomly shuffle leaf annotations and proceed with phylotype selection using the same procedure and selection criteria used for the original data. This shuffling procedure is repeated a number of times, typically 100 or 1,000, and p-values are computed. The implicit null hypothesis is that the annotations are randomly associated with the leaves of the tree. The p-value corresponds to the fraction of shuffled data sets in which one finds a phylotype with the annotation being evaluated and at least as large a criterion value as the observed phylotype. *Size* is the most discriminating criterion, and the use of the shuffling procedure is especially relevant with this criterion.

The Web interface (www.phylotype.org) divides analyses into five steps: *Input*, *Tree*, *Annotation*, *Analysis* and *Output*. The *Input* step enables to copy/paste or upload a phylogenetic tree and its annotations. The *Tree* step is optional, it integrates several rooting methods (*e.g.*, midpoint, outgroup). The *Annotation* step is also optional. It displays the annotation variables with all their possible values and respective frequencies among taxa. It also allows for logical combinations of annotations. For instance, an 'OR' connector allows for the aggregation of values (*e.g.*, making global 'regions' from 'countries' annotations). Other operators are available, such as logical 'AND', duplication, deletion, etc. The *Analysis* step corresponds to the PhyloType analysis itself. The user selects the criteria to be used for identifying phylotypes. Each criterion has a default threshold, which can be modified to be more or less stringent. An annotation variable must be selected and, for this variable, a set of annotations to study is chosen. Lastly, the user can decide to perform shuffling, define the number of iterations, and provide a p-value for phylotype selection using the *Size* criterion. PhyloType results are first summarized in an overview table that displays the number of phylotypes for each annotation value and various statistics. PhyloType analyses are fast, and it is easy to tune the parameter settings to explore the data with respect to this overview table. The final step (*Output*) gives access to the detailed results of PhyloType. A first table lists all selected phylotypes with corresponding statistics and taxa. Graphic outputs are available, such as phylogenies with color-encoded phylotypes or phylotype maps. All the input/output are available for download: original and rooted trees with node identifiers; annotations (primary and combined); overview and detailed phylotype tables; ready to print trees and map graphics in various formats.

## 3 Worldwide evolutionary history of HIV-1 subtype C

In order to demonstrate the abilities and efficiency of PhyloType, we applied the software to a large amount of data related to the HIV-1 subtype C epidemic on a worldwide scale using the *pol* sequences of [4], corresponding to all subtype C *pol* data available in the Los Alamos HIV database at the time of this study. We grouped countries in major areas in order to have enough strains for each group and to be able to extract synthetic and significant information. The results are reported in the following figure:




**Phylotype map of the worldwide study of HIV-1 subtype C**. Some of the phylotypes (colored in red) have indirect origin; for example, 789:Eastern Europe, with Southern Africa annotation along the path to 1:Central Africa.

Eighteen phylotypes are found covering 58% of the sequences:

- The analysis suggests that the epicenter of the HIV-1C epidemic is located in southern Central Africa (no. 1). This annotation includes strains collected in Zambia and Democratic Republic of Congo (DRC). The latter was previously identified as being the epicenter of viruses belonging to the pandemic group M of HIV-1 [5].

- The virus spread from Central Africa across the entire African continent, that is, directly into Southern African countries (no. 4578), South Africa (nos. 964 and 1445), East Africa (no. 4844), West Africa (no. 389) and Horn of Africa (no. 183), or indirectly to the Horn of Africa (nos. 4913, 4974 and 5063) and Southern African countries (no. 1488), passing through East and South Africa, respectively. However, the African phylotypes altogether cover only ~50% of the African strains: some regions are well-covered (*e.g.*, Central Africa, >70%), some are in between (*e.g.*, South and West Africa, ~50%), while the Southern African countries (Botswana, Mozambique, Malawi…) have a low coverage (~12%), likely due to their passing position between Central Africa and South Africa, with numerous introductions and bi-directional exchanges.
- The analysis suggests that the HIV-1 subtype C epidemic propagated from East Africa (no. 4844) towards South America (no. 5204, coverage of 99%), a link already identified upon several occasions (*e.g.* [6]).
- Our analysis indicates that the two Asian phylotypes (nos. 2866 and 3321, total coverage of 96%) originated in southern Central Africa (phylotype no. 1), which differs from the origin predicted for India (97% of Asian strains) by [7], suggesting South Africa instead. In this case as with South America, we see that a few major introductions suffice to explain the history of most of the current epidemic.
- This contrast with Europe, where (93%) of strains are not included in any phylotype, thus confirming multiple introductions and complex transmission chains of HIV-1C in Europe, as already pointed out in numerous studies. Only two low-sized (but significant) phylotypes are found. The Western Europe phylotype (no. 4310, descendant of Central Africa) only includes strains originating from Belgium, a country with historical and economic links with DRC. The Eastern Europe phylotype (no. 789) also contains strains from Romania exclusively.

## 4 Conclusion

To summarize results regarding the HIV-1 subtype C pandemic: PhyloType recovers a number of already identified transmission chains (*e.g.* East Africa to South America), contradicts a few others (*e.g.* origin of Indian epidemic) and suggests some new routes (*e.g.* two different geographical origins of the Horn of Africa epidemic). PhyloType results are obtained rapidly, in a matter of minutes, through a user-friendly Web interface. Therefore, we believe that PhyloType software has the potential to be useful for exploring and interpreting large virus phylogenies (HIV, Hepatitis B and so on), which are available today and should become the norm in the near future. The program's speed should be a major asset in surveillance tasks, a type of application we are currently working on. Moreover, we are using PhyloType with different biological models and aims, namely the geographical origin of the genus Coffea and the ecological speciation process among insects. PhyloType could also be used for radically different purposes, such as exploring large protein families associated with differentiated or specialized functions.

## References

[1] F. Chevenet, M. Jung, M. Peeters, T. de Oliveira and O. Gascuel, Searching for virus phylotypes. *Bioinformatics*, 29:561-70, 2013.

[2] D. Maddison and WP. Maddison, *MacClade 4: Analysis of Phylogeny and Character Evolution MacClade 4*, Sinauer Associates, Sunderland, MA, 2003.

[3] RG. Wallace, H. Hodac, RH. Lathrop and WM. Fitch, A statistical phylogeography of Influenza A H5N1. *Proc Natl Acad Sci USA.*, 104:4473-4478, 2007.

[4] M. Jung, N. Leye, N. Vidal, D. Fargette, H. Diop, C. Toure-Kane, O. Gascuel and M. Peeters, The origin and evolutionary history of HIV-1 subtype C in Senegal. *PLoS One*, 7:e33579, 2012.

[5] BF. Keele, F. Van Heuverswyn, Y. Li, E. Bailes, J. Takehisa, ML. Santiago, F. Bibollet-Ruche, Y. Chen, LV. Wain, F. Liegeois, S. Loul, EM. Ngole, Y. Bienvenue, E. Delaporte, JF. Brookfield, PM. Sharp, GM. Shaw, M. Peeters and BH. Hahn, Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*, 313:523-526, 2006.

[6] NM. Véras, RR. Gray, LF. Brígido, R. Rodrigues and M. Salemi, High-resolution phylogenetics and phylogeography of human immunodeficiency virus type 1 subtype C epidemic in South America. *J Gen Virol.*, 92:1698-1709, 2011.

[7] C. Shen, J. Craigo, M. Ding, Y. Chen and P. Gupta, Origin and dynamics of HIV-1 subtype C infection in India. *PLoS One*, 6:e25956, 2011.