Improved sensitivity and reliability of anchor based genome alignment

Raluca Uricaru, Célia Michotey, Laurent Noé, Helene H. Chiapello, Eric Rivals

Journées Ouvertes en Biologie Informatique Mathématiques

Nantes, Cité Internationale des Congrès,
9 au 11 Juin 2009

Avant-propos

Comme toutes les autres éditions depuis 2000, l'édition 2009 de JOBIM est avant tout une rencontre : celle de trois disciplines qui se donnent constamment rendez-vous depuis 10 ans, celle de plus de 300 chercheurs passionnés de recherches interdisciplinaires, celle de modestes essais et de grands résultats mis côte à côte dans un même ouvroir de conquêtes scientifiques.

Cette année, JOBIM doit beaucoup au groupe de travail de Bio-Informatique Ligérienne (BIL), un projet régional dont les participants sont les organisateurs de JOBIM. Regroupant des chercheurs des places nantaise et angevine, BIL est synonyme d'une recherche en bioinformatique qui a trouvé sa place et ses repères dans les Pays de la Loire. L'organisation de JOBIM est l'un, et pas le moindre, des effets mobilisateurs de BIL.

Nous avons reçu 32 soumissions comme communications longues et 49 soumissions comme posters. Le Comité de Programme aidé par quelques relecteurs additionnels a eu comme objectif d'identifier les soumissions les plus appropriées pour un exposé à JOBIM en terme (entre autres) de caractère novateur, de qualité des résultats, de niveau de présentation. Il a ainsi sélectionné 17 communications longues et 10 posters avec communication courte. Les soumissions non sélectionnées pour une présentation orale donnent lieu à 54 posters.

Nous avons le plaisir de compter parmi les participants et orateurs six conférenciers invités, dont les conférences ouvrent les six demi- journées thématiques. Le premier jour, Christian Gautier de l'université de Lyon I discourra sur la sélection et l'aléatoire en évolution, puis Jens Stoye de l'université de Bielefeld en Allemagne présentera des méthodes de métagénomique basée sur du séquençage haut-débit dans session algorithmique. Le second jour, Philipp Bucher, membre du Swiss Institute for Experimental Cancer Research (ISREC) et du Swiss Institute of Bioinformatics (SIB) nous montrera comment les expériences d'interaction entre protéine-ADN par séquençage (ChIP-seq) éclairent le fonctionnement des facteurs de transcription, tandis que Jean-Pierre Rousset de l'université d'Orsay exposera une des subtilités de la traduction d'ARN en protéine qu'est le recodage. Durant la troisième journée de JOBIM, Denis Thieffry de l'université de la Méditerranée à Marseille développera les modèles des réseaux de régulation et enfin, Gilbert Deléage de l'université Claude Bernard de Lyon parlera des programmes d'analyse de séquences et structures 3D de protéines et de leur intégration. Grâce à eux, la bioinformatique apparaîtra à JOBIM sous son meilleur jour, fascinante et pleine de promesses tenues.

JOBIM existe et se perpétue grâce aux chercheurs, réunis depuis quelques années dans la Société Française de BioInformatique (SFBI), et aussi grâce aux contributions financières de nos partenaires institutionnels, territoriaux ou industriels. Nous tenons à remercier chaleureusement les membres de BIL, la SFBI, les membres du comité de programme et les relecteurs additionnels, les six conférenciers invités qui nous font le plaisir de se joindre à nous. Nous exprimons toute notre gratitude aux partenaires de cette aventure : la Région Pays de la Loire, les universités de Nantes et d'Angers, l'INSERM, la fédération de recherches AtlanSTIC, le GDR de Bioinformatique Moléculaire, ReNaBi, l'unité INSERM U915, le LINA, le LERIA, Biogenouest, les départements CEPIA et MIA de l'INRA, Polytech'Nantes, l'entreprise Genomatix Software GmbH, ainsi qu'à la Ville de Nantes et la communauté urbaine de Nantes pour son accueil chaleureux.

Eric Rivals et Irena Rusu

Comité de programme

Eric Rivals (LIRMM, Montpellier), **Irena Rusu** (LINA-ComBi, Nantes)

Sébastien	Aubourg	Christine	Gaspin
Gregory	Batt	Robin	Gras
Séverine	Bérard	Yann	Guermeur
Vincent	Berry	Stéphane	Guindon
Michael	Blum	Patricia	Hernandez
Anne-Claude	Camproux	Pascal	Hingamp
Alessandra	Carbone	Hervé	Isambert
Cédric	Chauve	Fabien	Jourdan
Hélène	Chiapello	Gregory	Kucherov
Thérèse	Commes	Nicolas	Lartillot
François	Coste	Laurent	Noé
Miklos	Csuros	Elisabeth	Pécou
Antoine	Danchin	Adrien	Richard
Florence	d'Alche-Buc	Hughes	Richard
Alexandre	de Brevern	Hugues	Roest Crollius
Hidde	de Jong	Sophie	Schbath
Philippe	Derreumaux	Benno	Schwikowski
Gilles	Didier	Eric	Tannier
Thomas	Faraut	Pascal	Touzet
Guillaume	Fertin	Jean-Philippe	Vert
Christine	Froidevaux	Stéphane	Vialette

Nous remercions l'ensemble des relecteurs additionnels qui ont participé à l'évaluation des soumissions de JOBIM 2009. *Many thanks to the additional reviewers which participated to the assessment of submissions to JOBIM 2009*.

Relecteurs additionnels

Christophe	Ambroise	Dimitri	Gilis
Jean-Christophe	Avarre	Marco	CosentinoLagomarsino
Julie	Bernauer	Frédéric	Lemoine
Hugues	Berry	Oded	Maler
Martine	Boccara	Pierre	Peterlongo
Pascal	Bochet	Yann	Ponty
Alexander	Bockmayr	Bastien	Rance
Adrien	Bonneu	Elisabeth	Rémy
Anthony	Boureux	Hervé	Rey
Gilles	Caraux	Adrien	Richard
Sarah	Cohen Boulakia	Stéphanie	Sidibe-Bocs
Ludovic	Cottret	Fabienne	Thomarat
Georges	Czaplicki	Jean-Stéphane	Varré

Comité d'organisation

Rémi Houlgatte (INSERM U915, Nantes), **Jean-Michel Richer** (LERIA, Angers)

Sébastien	Angibaud
Daniel	Baron
Stéphane	Bezieau
Audrey	Bihouée
Jérémie	Bourdon
Henri	Briand
Solenne	Carat
Catherine	Chevalier
Freddy	Cliquet
Olivier	Collin
Audrey	Donnart
Raïssa	du Fretay
Emeric	Dubois
Béatrice	Duval
Damien	Eveillard
Guillaume	Fertin
Eric	Fonteneau
Delphine	Guillotin
Isabelle	Guisle
Jin-Kao	Hao
Yannick	Jacques
Pascale	Kuntz
Philippe	Leray
Virginie	Lollier
Morgan	Magnin
Yves	Malthiery
Mylène	Maurin
Hela	Memni
Raphaël	Mourad
Hoai-Tuong	Nguyen
Loic	Pauleve
Fabien	Picarougne
Mahatsangy	Raharijaona
Gérard	Ramstein
Olivier	Roux
Irena	Rusu
Frédérique	Savagnier
Jean-Jacques	Schott
Christine	Sinoquet
Dominique	Tessier
Raluca	Teusan

Table des matières

Conférenciers invités

Articles

Présentations Courtes

Posters

Conférenciers invités

Selection and randomness in evolution or selection of randomness in evolution ?

par **Christian Gautier**, UMR CNRS 5558, Université de Lyon

Biological evolution is based upon two processes, one generates genomic diversity (the mutational process) and one acts as a filter on this diversity (the selection process). For about fifty years biologists have addressed the question of the relative role of each of these processes on the patterning of genomes. If one "type" of mutation is particularly frequent, the filter has a higher probability to retain some of them but if a mutation is well "adapted" to the filter, even if it appears unfrequently, it has also a great probability to be retained. It is therefore mathematically difficult to discriminate between "mutational bias" and selective advantage.

Knowledge on the mutational process is quickly increasing with the accumulation of results on the cellular functioning. This process appears more and more complex and is also submitted to biodiversity : all organisms do have not the same mutational bias. Ecology (in a broad sense) has brought many recent results on selection, adaptation, genetic random drift and their relationships even in complexly structured populations. However a synthetic view of these two fields is difficult due to the very different scales both at the level of time (a few days to millions of years) and of biological organisation (molecules v.s. populations and communities). Moreover the results belong to very different biological fields and different biological communities (now in two different CNRS institutes !)

Being neither a cellular biologist nor an ecologist I do not claim that I shall present a new synthetic view of evolution ! The very modest aim of this talk is to explore the potentiality of bioinformatics, and particularly modeling, to help integrate population genetics, phylogeny, molecular evolution, cellular and molecular biology.

Computational Short Read Metagenomics

par **Jens Stoye**, Université de Bielefeld, Allemagne

Metagenomics is a new field of research on metagenomes, where natural microbial communities are studied. The new sequencing techniques like 454 or Solexa-Illumina sequencing promise new possibilities as they are able to produce huge amounts of data in much shorter time and with less efforts and costs than the traditional Sanger sequencing. But the data produced comes in even shorter reads (35-50 base pairs with Solexa-Illumina, 100-300 basepairs with 454 sequencing). CARMA [1] is a new pipeline for the characterization of the species composition and the genetic potential of microbial samples using 454-sequenced reads. The species composition can be described by classifying the reads into the taxonomic groups of organisms they most likely stem from. By assigning the taxonomic origins to the reads, a profile is constructed which characterizes the taxonomic composition of the corresponding community. The CARMA pipeline has already been successfully applied to 454-sequenced communities [2,3] including the characterization of a plasmid sample isolated from a wastewater treatment plant [4].

Using samples from a biogas plant we examined the applicability of this approach for the ultra-short Solexa-Illumina reads by comparing the results with those obtained by the 454-sequenced sample [5,6]. Our results using 77 million 50 bp-reads revealed that this approach indeed produces consistent results. Most diffences we have found are in the taxa of higher order, e.g. in the species level, and in general for species with a very low presence.

In order to apply CARMA to high-throughput sequencing data, we had to improve the accuracy and speed of our method in various ways : A preprocessing assembly phase using an adapted q-gram index [7] ; adaptation of the pipeline to take the information of mated reads into account to ”increase” read length ; modification of the amino acid sequence distance function for the construction of the phylogenetic tree ; and implementation of a protein-q-gram index over a multiple alignment for the read-against-Pfam protein family matching.

Références :

[1] L. Krause, N.N. Diaz, A. Goesmann, S. Kelley, T.W. Nattkemper, F. Rohwer, R.A. Edwards, J. Stoye (2008) Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. 36(7) :2230-2239.

[2] E.A. Dinsdale, O. Pantos, S. Smriga, R.A. Edwards, F. Angly, L. Wegley, M. Hatay, D. Hall, E. Brown, M. Haynes, L. Krause, E. Sala, S.A. Sandin, R. Vega Thurber, B.L. Willis, F. Azam, N. Knowlton, F. Rohwer (2008) Microbial Ecology of Four Coral Atolls in the Northern Line Islands. PLoS ONE 3(2) :e1584.

[3] S.A. Sandin, J.E. Smith, E.E. DeMartini, E.A. Dinsdale, S.D. Donner, A.M. Friedlander, T. Konotchick, M. Malay, J.E. Maragos, D. Obura, O. Pantos, G. Paulay, M. Richie, F. Rohwer, R.E. Schroeder, S. Walsh, J.B.C. Jackson, N. Knowlton, E. Sala (2008) Baselines and Degradation of Coral Reefs in the Northern Line Islands. PLoS ONE, 3(2) :e1548.

[4] A. Schlüter, L. Krause, R. Szczepanowski, A. Goesmann, A. Pühler (2008) Genetic diversity and composition of a plasmid metagenome from a wastewater treatment plant. J. Biotechnol. 136(1-2) :65-76.

[5] L. Krause, N.N. Diaz, R.A. Edwards, K.-H. Gartemann, H. Krömeke, H. Neuweger, A. Pühler, K.J. Runte, A. Schlüter, J. Stoye, R. Szczepanowski, A. Tauch, A. Goesmann (2008) Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor. J. Biotechnol. 136(1-2), 91-101.

[6] A. Schlüter, T. Bekel, N.N. Diaz, M. Dondrup, R. Eichenlaub, K.-H. Gartemann, I. Krahn, L. Krause, H. Krömeke, O. Kruse, J.H. Mussgnug, H. Neuweger, K. Niehaus, A. Pühler, K.J. Runte, R. Szczepanowski, A. Tauch, A. Tilker, P. Viehöver, A. Goesmann (2008) The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. J. Biotechnol. 136(1-2) :77-90.

[7] K. Rasmussen, J. Stoye, E.W. Myers (2006) Efficient q-Gram Filters for Finding All epsilon-Matches over a Given Length. J. Comp. Biol. 13(2) :296-308.

What directs a transcription factor to its target sites ? New insights from ChIP-Seq data analysis

par **Philipp Bucher**, Swiss Institute of Bioinformatics, Université de Lausanne

Chromatin immunoprecipitation combined with ultra-high throughput sequencing (ChIP-Seq) is revolutionizing research on gene regulation. This new technique allows for genome-wide mapping of in vivo occupied transcription factor binding sites in a quantitative manner at near-base pair resolution. Data on new transcription factors are released to the public almost every week. In most cases, computational analysis of the ChIP-Seq data has confirmed that transcription factors recognize the same DNA sequence motifs in vivo and in vitro. However, it has also become clear that only a small fraction of high-affinity sites to a transcription factors are occupied in vivo in a given cell type. Which are the additional determinants causing a transcription factors to bind to some target sites but not to others ? This is the key question addressed in this talk. Results from my group show that in vivo transcription factor binding sites are distinct from sequence-wise identical non- occupied sites in terms of cross-species conservation pattern, chromatin conformation, and flanking sequence motif content.

Le recodage : Qu'est-ce que c'est ? A quoi ça sert ? Comment ça marche ? Et la bioinformatique dans tout ça ?

par **Jean-Pierre Rousset**, Institut de Génétique et Microbiologie, Université Paris-Sud, Orsay

Identifiés initialement chez les virus à ARN et les transposons, les événements de décodage non conventionnel de l'information génétique (ou « recodage ») ont depuis été observés chez une grande variété d'organismes. Ils consistent en une modification locale des règles standards de décodage par un ensemble de signaux, séquences et structures, sur l'ARN messager, capables de perturber les acteurs normaux de la synthèse protéique. Ainsi, la machinerie de traduction (ribosome, ARN de transferts, facteurs variés) va parfois incorporer un acide aminé à l'endroit d'un codon de terminaison, ou changer de cadre de lecture en permettant, dans la plupart des cas, d'éviter un codon de terminaison. Ces événements ont une efficacité de l'ordre de 2 à 50% et entraînent donc la synthèse de deux polypeptides à partir d'un même ARNm, l'un portant un domaine fonctionnel supplémentaire par rapport à celui prédit en utilisant le code génétique « universel ».

Le recodage est utilisé par les cellules et de nombreux virus pour contrôler, quantitativement ou qualitativement, l'expression de certains gènes. A titre d'exemple, la télomérase de plusieurs levures, l'ADN polymérase de nombreuses bactéries et la transcriptase inverse d'une multitude de virus, dont le VIH, nécessitent un événement de recodage pour être exprimés. Récemment, notre équipe a montré qu'un gène contrôlé par un événement de décalage de cadre de lecture en +1 est la première cible identifiée d'un prion de la levure Saccharomyces cerevisiae.

L'étude de ces événements est intéressante à deux titres. D'une part, elle constitue une approche originale pour comprendre les mécanismes de la traduction, un peu sur le même principe que le mutant renseigne sur le fonctionnement normal de la cellule. Le recodage permet ainsi d'identifier les éléments de la machinerie traductionnelle qui sont impliqués dans la reconnaissance d'un codon de terminaison ou le maintien du cadre de lecture. D'autre part, l'étude de l'interaction des signaux mis en jeu, séquences et structures, avec la machinerie de traduction permet de caractériser de nouveaux éléments fonctionnels dont le rôle n'était pas soupçonné.

Finalement, les événements de recodage posent deux types de défis aux bioinformaticiens : i) comment identifier des objets biologiquement pertinents à partir d'une connaissance parfois très imparfaite des signaux qui les caractérisent ; ii) comment reconnaître les structures de l'ARNm impliquées dans le recodage. L'accumulation des séquences de génomes complets permet de penser que la génomique comparative aidera, dans un avenir proche, à résoudre le premier problème. C'est plus probablement par des approches d'informatique « dure » que l'on peut espérer que viendront les prochaines avancées pour le deuxième aspect qui nécessite de comparer ou d'analyser la conservation de structures.

Tackling regulatory networks : from biological models to theorems, and vice-versa

par **Denis Thieffry**, INSERM U928, Université de la Méditerranée, INRIA Paris-Rocquencourt

At Luminy, biologists, computer scientists and mathematicians are collaborating on the discrete, dynamical modelling of biological regulatory networks since several years. This talk will describe our modes of collaboration and overview our main results, from the formulation of mathematical theorems to the design of Petri nets, and from software development to the specification and analysis logical models for the control of cell proliferation and differentiation processes.

Références :

[1] Chaouiya C, Remy E, Mossé B, Thieffry D (2003). Qualitative Analysis of Regulatory Graphs : A Computational Tool Based on a Discrete Formal Framework. Lect Notes Control Info Sci 294 : 119-126.

[2] Chaouiya C, Remy E, Ruet P, Thieffry D (2004). Qualitative Modelling of Genetic Networks : From Logical Regulatory Graphs to Standard Petri Nets. Lect Notes Comput Sci 3099 : 137-56.

[3] Chaouiya C, Remy E, Thieffry D (2008). Petri net modelling of biological regulatory networks. J Disc Algo 6 : 165-77.

[4] Fauré A, Naldi A, Chaouiya C, Thieffry D (2006). Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. Bioinformatics 22 : e124-31.

[5] González AG, Naldi A, Sánchez L, Thieffry D, Chaouiya C (2006). GINsim : a software suite for the qualitative modelling, simulation and analysis of regulatory networks. Biosystems 84 : 91-100.

[6] González AG, Chaouiya C, Thieffry D (2008). Qualitative dynamical modelling of the formation of the anterior-posterior compartment boundary in the Drosophila wing imaginal disc. Bioinformatics 24 : i234-40.

[7] Gonzalez AG, Chaouiya C, Thieffry D (2006). Dynamical analysis of the regulatory network defining the dorsal-ventral boundary of the Drosophila wing imaginal disc. Genetics 174 : 1625-34.

[8] Remy E, Mosse B, Chaouiya C, Thieffry D (2003). Discrete dynamics of regulatory feedback circuits. Bioinformatics 10 : ii172-8.

[9] Remy E, Ruet P, Thieffry D (2006). Positive or negative regulatory circuit inference from multilevel dynamics. Lect Notes Control Info Sci 341 : 263-70.

[10] Naldi A, Thieffry D, Chaouiya C (2007). Decision diagrams for the representation and analysis of logical models of genetic networks. Lect Notes Comput Sci 4695 : 233-47.

[11] Remy E, Ruet P, Thieffry D (2008). Graphic requirements for multistability and attractive Cycles in a Boolean dynamical framework. Adv App Math 41 : 335-50.

[12] Simão E, Remy E, Thieffry D, Chaouiya (2005). Qualitative Modelling of Regulated Metabolic Pathways : Application to the Tryptophan Biosynthesis in E. Coli. Bioinformatics 21 : ii190-6.

Méthodes d'analyse de séquences et de structures 3D de protéines et leur intégration au sein de Webiciels

par **Gilbert Deléage**, IBCP, UMR 5086, CNRS, Université de Lyon

La fonction biologique d'une protéine est intimement liée à sa structure 3D et à ses interactions. Depuis de nombreuses années, nous développons au sein de notre équipe à l'IBCP de Lyon (`http://pbil.ibcp.fr`) , des outils intégrés d'analyse de séquences de protéines (serveur NPS@ : http ://npsa-pbil.ibcp.fr), des outils de modélisation moléculaire par homologie en particulier à faible taux d'identité en utilisant les prédictions de structures secondaires comme Geno3D (`http://geno3d-pbil.ibcp.fr`). Cet outil de modélisation s'est récemment enrichi du pipeline d'annotation structurale MAGOS, en collaboration avec le groupe d'O. Poch à l'IGBMC, et d'un système de modélisation moléculaire à haut débit permettant le traitement au niveau d'un protéome entier. Un système de gestion des modèles 3D a été développé sous la forme d'une base de données afin de faciliter les mises à jour et les requêtes concernant l'exploitation de ces modèles MODEOME3D. Ces outils sont intégrés au sein du serveur PIG (Protein InvestiGator, `http://pig-pbil.ibcp.fr`). Au niveau des développements méthodologiques, l'outil SuMo (http ://sumo-pbil.ibcp.fr) permet de comparer des structures 3D de protéines en s'affranchissant du repliement 3D des protéines. Il utilise une description des protéines de la PDB sous forme de graphes de triplets de groupements physico-chimiques connectés. Cet outil permet de réaliser une annotation fonctionnelle des structures 3D sans fonction connue issues des grands programmes de génomique structurale. Enfin, des outils de validation de modèles 3D en utilisant des approches expérimentales de spectrométrie de masse ont été récemment mis au point (`http://proteomics-pbil.ibcp.fr`). Une revue de l'ensemble des méthodes et outils développés sera effectuée à travers des exemples d'applications biologiques.

Articles

Single-nucleotide substitution rates increase during the replication S phase of the human genome

Chun-Long Chen[1], Aurélien Rappailles[2], Lauranne Duquenne[1,4], Maxime Huvet[1,5], Guillaume Guilbaud[2], Benjamin Audit[3], Yves d'Aubenton-Carafa1, Alain Arneodo[3], Olivier Hyrien[2] & Claude Thermes[1]

[1]Centre de Génétique Moléculaire (CNRS), Allée de la Terrasse, 91198 Gif-sur-Yvette, France
chen@cgm.cnrs-gif.fr, daubenton@cgm.cnrs-gif.fr, thermes@cgm.cnrs-gif.fr
[2]Ecole Normale Supérieure de Paris, 46 rue d'Ulm, 75005 Paris
rappaill@biologie.ens.fr, guilbaud@biologie.ens.fr, hyrien@biologie.ens.fr
[3]Laboratoire Joliot Curie et Laboratoire de Physique, Ecole Normale Supérieure de Lyon, 69364 Lyon
baudit@ens-lyon.fr, alain.arneodo@ens-lyon.fr
[4]Present address: UMR CNRS 5558, LBBE, UCB Lyon1, 43 bd du 11 Novembre 1918, 69622 Villeurbanne
duquenne@biomserv.univ-lyon1.fr
[5]Present address: Imperial College London, South Kensington Campus, London SW7 2AZ
m.huvet@imperial.ac.uk

Abstract: *Naturally occurring mutations in mammalian genomes play a key role in evolution and genetic disease but their causes are still poorly understood. In particular, nucleotide substitutions occur at strongly variable rates along genomes and it is essential to unravel the mechanisms responsible of these fluctuations. A number of evolutionary studies have exhibited complex correlations between substitution rates and parameters like regional or local nucleotide composition, crossover rate or distance to telomeres [1-5]. Here, we study the role of replication on neutral substitution rates in the human genome. Using replication timing data determined by massive sequencing of replicating strands, we show that all non-CpG substitution rates correlate with timing: they are minimum in early replicating regions and increase to maximum values in late regions. These correlations are still observed after controlling for nucleotide composition, cross-over rate and distance to telomeres. These data demonstrate for the first time that replication timing plays a key role in shaping the profile of mutations along the genome.*

Keywords: human genome, nucleotide substitutions, replication timing.

1 Introduction

Mutations are known to occur very heterogeneously along genomes but despite numerous studies the causes of these fluctuations remain relatively unknown. Fundamental questions like the role of replication-associated processes in these fluctuations remain to be addressed. During the last two decades numerous works exhibited an increasing complexity of mutation patterns. Substitutions depend on large scale nucleotide composition as in the case of the GC content [6] or on the nucleotide flanking the mutated site as for the CpG dinucleotides [7] as well as on other nucleotide contexts [8,9]. Recombination correlates with global substitution rate [10] and

with the ratio of W (A or T)→S (G or C) to S→W substitution rates [11], this correlation likely resulting from biased gene conversion [1-3]. It was hypothesized in precursor studies that changes of nucleotide pools during replication would be responsible of mutation rate fluctuations along the genome [12] but lack of replication timing data did not allow testing this possibility. In this work, we examine the role of replication on substitution rates in the human genome. Replication timing was determined along the genome and nucleotide substitutions rates were tabulated in the human lineage since its divergence with chimpanzee. Analysis of these data show that substitution rates are strongly dependent on the timing values, exhibiting an important role of replication on genome evolution.

2 Results

In order to determine the fluctuations of substitution rates during replication, the replication timing profile of the human genome was determined by immunoprecipitation of nascent replicating BrdU-labeled DNA fragments extracted from HeLa cells sorted at different periods of the S phase (Materials and Methods). Nucleotide substitutions were tabulated in the human lineage since its divergence with chimpanzee using macaque as an outgroup (Materials and Methods). The global substitution rate computed in 100 kbp DNA fragments shows strong variations with replication timing. Rate computed in non-coding sequences increases from minimum values (0.5%) in early replicating regions to 0.62% in late regions (Figure 1a). Several parameters are known to correlate with substitution rates, namely nucleotide composition, either regional GC content or local nucleotide context [9], crossover rate and distance to telomeres. Replication timing strongly correlates with GC content, GC-rich (resp. GC-poor) regions replicating mostly in early (resp. late) S phase [13]. We examined the correlations of the global rate with timing when controlling for local GC content, crossover rate or distance to telomeres (Figure 1b-d). In all cases the global rate increases regularly from early to late timing (rate increases in a proportion ranging between 30 and 50%). When further controlling simultaneously for all three parameters, the global substitution rate still strongly correlates with replication timing, within both repeated elements and non-repeated elements (data not shown). We also examined the correlation between human diversity and timing, using SNP data from several individual fully sequenced genomes. Similar increase of diversity with timing was observed for all genomes examined, further establishing the timing-dependence of substitution rates (data not shown). To disentangle the role of the various factors, multivariate regression analysis of substitution rates was performed with the 4 predictors, GC content, crossover rate, log of distance to telomere and replication timing. Replication timing is the best predictor of the variability explained by the model (28%) and explains 9.5% of overall variability in substitution rates.

We further investigated the impact of replication timing on individual substitution rates. All rates correlate positively with replication timing when controlling for GC content, crossover rate and distance to telomere (Figure 2). Interestingly, S→W and W→W rates show strong dependency on replication timing but weak dependency on the three other parameters (Figure 2a,c,e). Conversely, W→S and S→S rates show a much stronger dependency on these parameters (Figure 2b,d,f). Rates increase from early to late regions by a factor ranging from 1.3 for AT→GC to a maximum 1.8 for CG→AT.

It is known that substitution rates can depend on the flanking nucleotide context [9]. To eliminate possible context effects, we investigated two groups of transitions within a given flanking nucleotide context. We examined the two highest transition rates, TGT/ACA→TAT/ATA and TAT/ATA→TGT/ACA [9] and observed that they increase with replication timing as when flanking nucleotides are not fixed (Figure 3).

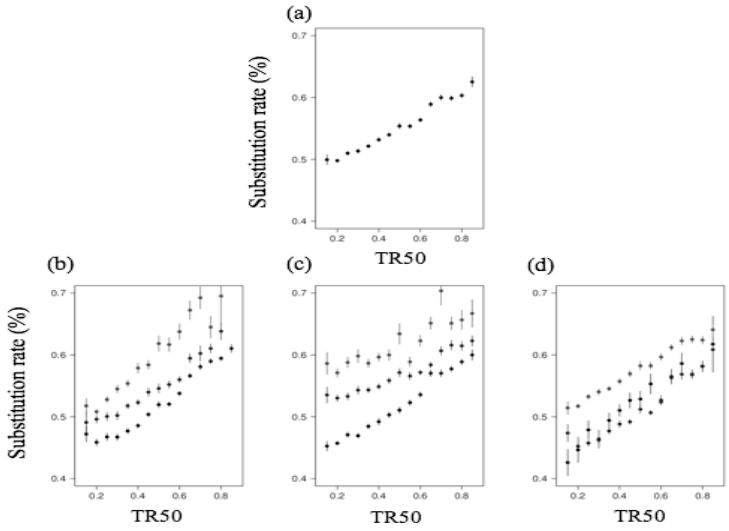


Figure 1. Correlations between total substitution rate and replication timing. The total rate of non-CpG substitutions is the rate of substitutions of all types per nucleotide computed by comparison of human and chimpanzee genome sequences (Materials and Methods). In abscissa, timing (TR50) values determined in 100 kbp windows (Materials and Methods); in ordinate, the mean value of total substitution rate ± SEM in percent. (a) Total rate as function of TR50. (b) Same as in (a) with control of the GC content of the analyzed segments; black, GC≤38%; dark grey, 38<GC≤43%; light grey, GC>43%. (c) Same as in (a) with control of the crossover rate; black, CO≤1 cM/Mb; dark grey, 1<CO≤3 cM/Mb; light grey, CO>3 cM/Mb. (d) Same as in (a) with control of the distance to telomeres of analyzed segments; black, DT>80 Mb; dark grey, 40<DT≤80 Mb; light grey, DT≤40 Mb.

.

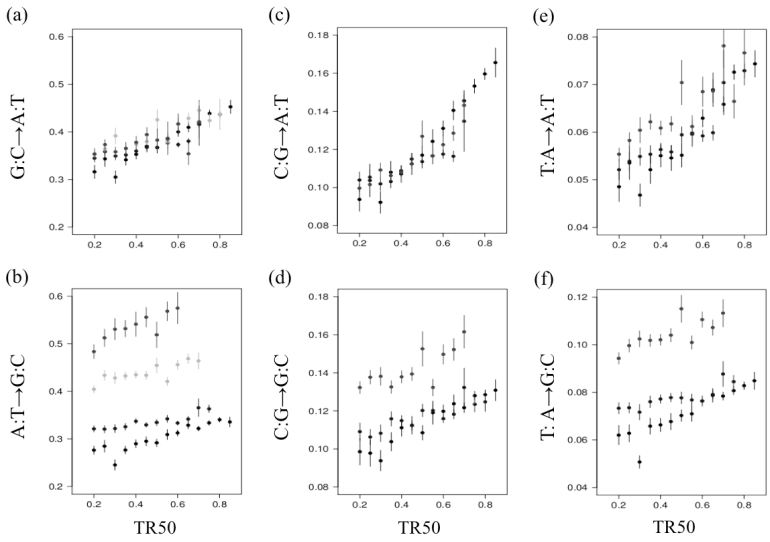


Figure 2. Variations of individual substitution rates with replication timing when controlling for GC content, crossover rate and distance to telomeres. Timing values and non-CpG substitution rates are determined as in Figure 1. Black, GC≤38%, CO≤1 cM/Mb, DT>80 Mb; dark grey, 38<GC≤43%, 1<CO≤3 cM/Mb, 40<DT≤80 Mb. For transitions (a,b), pale grey, 43<GC≤55%, CO>3 cM/Mb, DT≤40 Mb; light grey, GC>55%, CO>3 cM/Mb, DT≤40 Mb. For transversions (c-f), light grey, GC>43%, CO>3 cM/Mb, DT≤40 Mb.

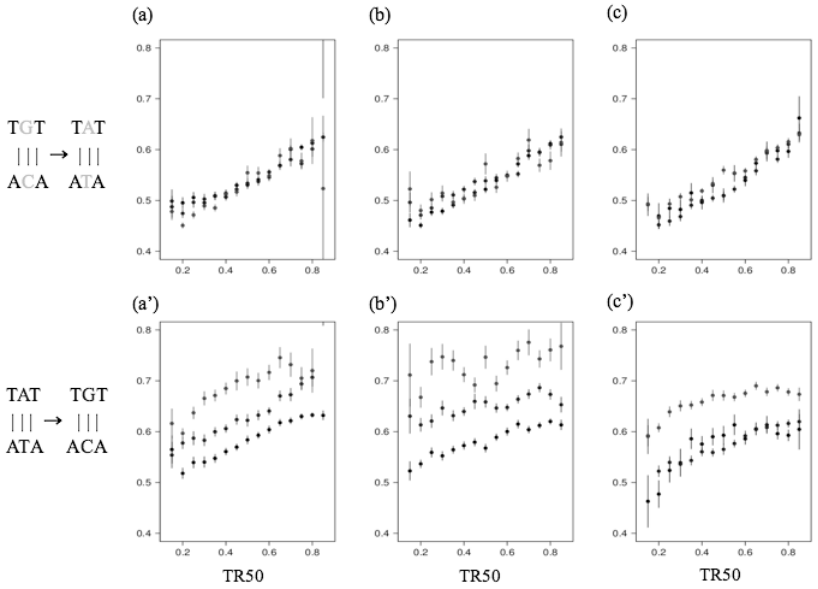


Figure 3. Variation of individual substitution rates with replication timing with control of flanking nucleotide context. Timing values and substitution rates are determined as in Figure 1. Two groups of transitions rates within given flanking nucleotide context as function of TR50 with control of the GC content (a,a'), black, GC<38%; dark grey, 38%<GC≤43%; light grey, GC>43%; with control of the crossover rate (b,b'), black, CO≤1 cM/Mb; blue, 1<CO≤3 cM/Mb; light grey, CO>3; with control of the distance to telomeres (c,c'), black, DT>80 Mb; dark grey, 40<DT≤80 Mb; light grey, DT≤40 Mb.

A recent analysis showed that overall substitution rate correlates with chromatin structure [14]. Rates are lower in genomic regions with open chromatin structure than in regions with closed chromatin structure. The authors proposed that this would result from lower accessibility of closed chromatin to repair machineries. Replication timing strongly correlates with chromatin structure, open (resp. closed) regions replicating mostly in early (resp. late) S phase [15], which could explain the correlation observed in the present work. When controlling for the local chromatin structure, the global substitution rate still increases with replication timing (data not shown) showing that the correlations between rates and timing unlikely result from changes of chromatin structure during the S phase.

3 Discussion

Analysis of substitutions that occurred in the human lineage since its divergence with chimpanzee show that replication timing has a major impact on neutral substitution rates. This result assumes that replication timing remained mostly constant since human-chimpanzee divergence, a property that likely results from previous observation that timing remained mostly constant between human and mouse [16]. Several hypotheses can be considered to explain our observations. A possibility is that the increase of substitution rates during the S phase reflects a corresponding increase of replication errors. During the cell cycle, the pools of dNTPs are finely tuned to ensure the onset of replication [17]. It has been observed that alteration of dNTP pools can enhance the rate of replication errors [18]. It is then possible that during the S phase, dNTP amounts undergo physiological modulations that would ultimately induce the observed increase of substitution rates. Alternatively, increase of substitution rates during the S phase could result from a corresponding decrease of DNA repair. Correction of replication errors require MMR [19]. MMR is active in all phases of the cell

cycle [20] but its activity is enhanced during the S phase compared to G1 and G2 [21]. Variations of MMR activity measured during the cell cycle show that G:T mismatches are repaired better than G:A (leading to larger proportion of GC→TA) [21]. It is possible that MMR activity is highest at the onset of the S phase and then decreases progressively to low values in late S phase, correspondingly to the G2 level. According to this model, the rate of GC→TA would increase significantly more than GC→AT during the course of the S phase, in agreement with our observations (Figure 3c,a). Increase of replication errors and decrease of repair activity during the S phase could also contribute simultaneously to the observed increase of substitution rates. Clearly, these are only some of several mechanisms that can be responsible of the observed correlations. However, the data demonstrate for the first time in mammals, an essential role of replication-associated mutation and repair mechanisms in the variations of the neutral mutation pattern along the genome. They open new roads to study these mechanisms and o quantify their effects.

4 Material and Methods

Massive sequencing of BrdU-labeled nascent replicated DNA. Isolation of BrdU-labelled nascent strands was adapted from [22]. Asynchronous Hela cells were pulse-labeled with 50 µM bromodeoxyuridin (BrdU) for 40 min ; cells were collected in four different periods of S phase by FACS, namely S1, S2, S3 and S4. Cells collected during the entire S phase were used as control. BrdU-labelled DNA was immunoprecipitated with BrdU antibody. Double stranded DNA was obtained from immunoprecipitated DNA samples by random priming. Resulting DNA was sequenced using Illumina Solexa sequencing device.

Calculation of replication timing. Replication timing of a genome region was estimated by the time (TR50) at which 50% of reads mapping in this region are obtained (as described in [23]). Briefly, at first, we calculated the enrichment of sequence read, D, for each sample collected in different time points over the control sample within a given window (e.g. 100kb in this study). The enrichment value of each sample corresponds to replication within each quarter of the S phase. TR50 was calculated by using a linear interpolation for 50% enrichment values. When D=0 for all 4 time points, we did not make a TR50 estimation.

Determination of substitution rates. Sequence alignments of *Homo sapiens* (hg18), *Pan troglodytes* (panTro2) and *Macaca mulatta* (rheMac2) and human genome annotations were retrieved from UCSC Genome Brower (http://genome.ucsc.edu). Coding regions were not considered in the analyses. Nucleotide substitutions were tabulated in the human lineage since its divergence with chimpanzee using macaque as outgroup. To minimize effects of alignment artifacts, only isolated substitutions defined as those flanked by sites identical in all three species were tabulated. Non-CpG substitution rates were calculated within non-overlapping 100 kb windows by dividing the number of substitution events of appropriate type by the number of potentially mutable sites that meet the same criteria. We also performed 4-ways alignments between human and chimpanzee, using macaque and pongo as outgroups, thus ensuring that multiple substitutions were neglected. Crossover rate data were retrieved from the HAPMAP project (www.hapmap.org). The crossover rate for a given window was computed as a weighted average of crossover rates in chromosomal regions overlapping with the corresponding window.

Acknowledgements

We thank the cell sorting facility of the Institut Jacques Monod UMR CNRS 7592, Universities Paris VII and VI (supported by the Région Ile-de-France). This work was supported by the Centre National de la Recherche Scientifique (CNRS), the Agence Nationale de la Recherche (NT05-3_41825) and grants from the Association pour la Recherche sur le Cancer, the Ligue Contre le Cancer (Comité de Paris) and the Fondation pour la Recherche Médicale to O.H.

References

[1] N. Galtier, G. Piganeau, D. Mouchiroud and L. Duret. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, 159:907-911,2001.

[2] A. Eyre-Walker and L.D. Hurst. The evolution of isochores. *Nat. Rev. Genet.*, 2:549-555.,2001.

[3] L. Duret and P.F. Arndt. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*, 4:e1000071,2008.

[4] K.J. Fryxell and E. Zuckerkandl. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol*, 17:1371-1383.,2000.

[5] N. Elango, S.H. Kim, E. Vigoda and S.V. Yi. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. *PLoS Comput Biol*, 4:e1000015,2008.

[6] R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520-562,2002.

[7] M. Ehrlich and R.Y. Wang. 5-Methylcytosine in eukaryotic DNA. *Science*, 212:1350-1357,1981.

[8] Z. Zhao and E. Boerwinkle. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res*, 12:1679-1686,2002.

[9] D.G. Hwang and P. Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A*, 101:13994-14001,2004.

[10] M.J. Lercher and L.D. Hurst. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet*, 18:337-340,2002.

[11] J. Meunier and L. Duret. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol*, 21:984-990,2004.

[12] K.H. Wolfe, P.M. Sharp and W.H. Li. Mutation rates differ among regions of the mammalian genome. *Nature*, 337:283-285.,1989.

[13] K. Woodfine, H. Fiegler, D.M. Beare, J.E. Collins, O.T. McCann, B.D. Young, S. Debernardi, R. Mott, I. Dunham and N.P. Carter. Replication timing of the human genome. *Hum. Mol. Genet.*, 13:191-202,2004.

[14] J.G. Prendergast, H. Campbell, N. Gilbert, M.G. Dunlop, W.A. Bickmore and C.A. Semple. Chromatin structure and evolution in the human genome. *BMC Evol Biol*, 7:72,2007.

[15] N. Gilbert, S. Boyle, H. Fiegler, K. Woodfine, N.P. Carter and W.A. Bickmore. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell*, 118:555-566,2004.

[16] S. Farkash-Amar, D. Lipson, A. Polten, A. Goren, C. Helmstetter, Z. Yakhini and I. Simon. Global organization of replication time zones of the mouse genome. *Genome research*, 18:1562-1570,2008.

[17] C.M. Hu and Z.F. Chang. Mitotic control of dTTP pool: a necessity or coincidence? *J Biomed Sci*, 14:491-497,2007.

[18] M. Meuth. The molecular basis of mutations induced by deoxyribonucleoside triphosphate pool imbalances in mammalian cells. *Exp Cell Res*, 181:305-316,1989.

[19] T.A. Kunkel and D.A. Erie. DNA mismatch repair. *Annu Rev Biochem*, 74:681-710,2005.

[20] A.G. Schroering, M.A. Edelbrock, T.J. Richards and K.J. Williams. The cell cycle and DNA mismatch repair. *Exp Cell Res*, 313:292-304,2007.

[21] M.A. Edelbrock, S. Kaliyaperumal and K.J. Williams. DNA mismatch repair efficiency and fidelity are elevated during DNA synthesis in human cells. *Mutat. Res.*:Epub ahead of print,2008.

[22] V. Azuara. Profiling of DNA replication timing in unsynchronized cell populations. *Nat Protoc*, 1:2171-2177,2006.

[23] Y. Jeon, S. Bekiranov, N. Karnani, P. Kapranov, S. Ghosh, D. MacAlpine, C. Lee, D.S. Hwang, T.R. Gingeras and A. Dutta. Temporal profile of replication of human chromosomes. *Proc Natl Acad Sci U S A*, 102:6419-6424,2005.

Counting patterns in degenerated sequences

Grégory Nuel[1]

MAP5, UMR 8145 CNRS, University Paris Descartes
5 rue des Saints-Peres, F-75006 Paris, France
gregory.nuel@parisdescartes.fr

Abstract: *In this paper, we propose a rigourous method to take into account the uncertainty of sequencing for biological sequences (DNA, Proteins). For example, this method allows to study the distribution of a pattern of interest in a degenerated sequence defined on the standard IUPAC DNA alphabet. We first introduce a Forward-Backward approach to compute the marginal distribution of the constrained sequence and use it both to perform a Expectation-Maximization estimation of parameters, as well as deriving a heterogeneous Markov distribution for the constrained sequence. This distribution is hence used along with known DFA-based pattern approaches to obtain the exact distribution of the pattern count under the constraints. As an illustration, we consider a EST dataset from the EMBL database. Despite the fact that only* 1% *of the position in this dataset are degenerated, we show that not taking into account these positions might lead to erroneous observations, further proving the interest of our approach.*

Keywords: Markov chains, Expectation-Maximization, Posterior distribution, Forward-Backward, Moment generating function, Deterministic finite automaton

1 Introduction

Biological sequences like DNA or proteins, are always obtained through a sequencing process which might produce some uncertainty. As a result, such sequences are usually written in a degenerated alphabet where some symbols may correspond to several possible letters. For example, the IUPAC [1] protein alphabet includes the following degenerated symbols: `X` for "any amino-acid", `Z` for "glutamic acid or glutamine", and `B` for "Aspatic acid or Asparagine". For DNA sequences, there is even more of such degenerated symbols which exhaustive list and meaning are given in Table 1 along with observed frequencies in several datasets from the EMBL database [2].

When counting patterns in such degenerated sequences, the question that naturally arise is: how to deal with degenerated positions ? Since most (usually 99%) of the positions are not degenerated, it is usually considered harmless to discard those degenerated positions in order to get an observation. Another option might be to simply ignore the problem by considering the degenerated alphabet as a standard alphabet. Finally, one might come up with some *ad hoc* counting rule like: "whenever the pattern might occurs I add one [1] to the observed count". However practical, all these solutions remain quite unsatisfactory from the statistician point of view. In this paper, we want to deal rigourously with this problem by introducing the distribution of sequences under the uncertainty of their sequencing, and then by using this distribution to study the "observed" number of occurrences of a pattern of interest.

[1] one might also think to add a fraction of one which correspond to the probability to see the adequate letter at the degenerated position.

symbol	meaning	est_pro_01	htg_pro_01	htc_fun_01	std_hum_21
A	Adenine	67459	1268408	1347782	1190205
C	Cytosine	53294	1706478	1444861	1031369
G	Guanine	54194	1719016	1325070	809651
T	Thymine	66139	1277939	1334061	1067933
U	Uracil	0	0	0	0
R	Purine (A or G)	13	0	7	39
Y	Pyrimidine (C, T, or U)	6	0	9	37
M	C or A	2	0	6	31
K	T, U, or G	6	0	5	30
W	T, U, or A	6	0	8	26
S	C or G	21	0	4	28
B	not A	0	0	0	0
D	not C	3	0	0	0
H	not G	0	0	1	0
V	not G, not U	0	0	0	0
N	any base	1792	115485	28165	19272

Table 1. Meaning and frequency of the IUPAC [1] DNA symbols in several files of the release 97 of the EMBL nucleotide sequence database [2]. Degenerated symbols (lowest part of the table) contribute to 0.5% to 1% of the data.

2 Constrained distribution

Let $X_1^\ell = X_1 \dots X_\ell$ be a order [2] $d \geqslant 1$ homogeneous Markov chain over the finite alphabet $\mathcal{A}$ such as $\nu\left(a_1^d\right) \stackrel{\text{def}}{=} \mathbb{P}\left(X_1^d = a_1^d\right)$ and $\pi(a_1^d, b) \stackrel{\text{def}}{=} \mathbb{P}(X_{i+d} = b | X_i^{i+d-1} = a_1^d)$ for all $a_1^d \stackrel{\text{def}}{=} a_1 \dots a_d \in \mathcal{A}^d$, $b \in \mathcal{A}$, and $1 \leqslant i \leqslant \ell - d$. Our objective is to compute the constrained distribution $\mathbb{P}\left(X_1^\ell | X_1^\ell \in \mathcal{X}_1^\ell\right)$ where all $\mathcal{X}_i \subset \mathcal{A}$ are the subset of the possible values taken by X_i and $\mathcal{X}_1^\ell \stackrel{\text{def}}{=} \mathcal{X}_1 \times \ldots \times \mathcal{X}_\ell$. If $\mathcal{X}_i = \{x_i\}$, $x_i \in \mathcal{A}$ for all i, then we get the degenerated distribution concentrated at x; if $\mathcal{X}_i = \mathcal{A}$ for all i, then we get the initial unconstrained distribution of X.

PROPOSITION 2.1 (FORWARD). *For all $x_1^\ell \in \mathcal{A}^\ell$ and $\forall i, 1 \leqslant i \leqslant \ell - d$ we define the* forward *quantity $F_i\left(x_i^{i+d-1}\right) \stackrel{\text{def}}{=} \mathbb{P}\left(X_i^{i+d-1} = x_i^{i+d-1}, X_1^{i+d-1} \in \mathcal{X}_1^{i+d-1}\right)$ which is computable by recurrence through:*

$$F_i\left(x_i^{i+d-1}\right) = \sum_{x_{i-1} \in \mathcal{X}_{i-1}} F_{i-1}\left(x_{i-1}^{i+d-2}\right) \pi\left(x_{i-1}^{i+d-2}, x_{i+d-1}\right) \quad \forall i, 2 \leqslant i \leqslant \ell - d + 1 \qquad (1)$$

with the initialization $F_1\left(x_1^d\right) = \nu\left(x_1^d\right) \mathbb{I}_{\mathcal{X}_1^d}(x_1^d)$ where $\mathbb{I}$ is the indicatrix function [3]. We then obtain that:

$$\mathbb{P}(X_1^\ell \in \mathcal{X}_1^\ell) = \sum_{x_{\ell-d}^\ell \in \mathcal{X}_{\ell-d}^\ell} F_{\ell-d}\left(x_{\ell-d}^{\ell-1}\right) \pi\left(x_{\ell-d}^{\ell-1}, x_\ell\right). \qquad (2)$$

[2] the particular degenerated case where $d = 0$ is left to the reader for the sake of simplicity.

[3] for any set E, subset $A \subset E$ and element $a \in E$, $\mathbb{I}_A(a) = 1$ if $a \in A$ and $\mathbb{I}_A(a) = 0$ otherwise.

Proof. We prove Equation (1) by simply rewritting $F_i(x_i^{i+d-1})$ as:

$$\begin{aligned} F_i\left(x_i^{i+d-1}\right) &= \sum_{x_{i-1}\in\mathcal{X}_{i-1}} \mathbb{P}(X_{i-1}^{i+d-1} = x_{i-1}^{i+d-1}, X_1^{i+d-1} \in \mathcal{X}_1^{i+d-1}) \\ &= \sum_{x_{i-1}\in\mathcal{X}_{i-1}} \underbrace{\mathbb{P}(X_{i-1}^{i+d-2} = x_{i-1}^{i+d-2}, X_1^{i+d-2} \in \mathcal{X}_1^{i+d-2})}_{F_{i-1}\left(x_{i-1}^{i+d-2}\right)} \\ &\quad \times \underbrace{\mathbb{P}(X_{i+d-1} = x_{i+d-1}, X_{i+d-1} \in \mathcal{X}_{i+d-1} | X_{i-1}^{i+d-2} = x_{i-1}^{i+d-2}, X_1^{i+d-2} \in \mathcal{X}_1^{i+d-2})}_{\pi\left(x_{i-1}^{i+d-2}, x_{i+d-1}\right)\mathbb{I}_{\mathcal{X}_{i+d-1}}(x_{i+d-1})}. \end{aligned}$$

The proof of Equation (2) is established in a similar manner. □

PROPOSITION 2.2 (BACKWARD). *For all $x_1^\ell \in \mathcal{A}^\ell$ and $\forall i, 1 \leqslant i \leqslant \ell - d$ we define the* backward *quantity $B_i\left(x_i^{i+d-1}\right) \stackrel{def}{=} \mathbb{P}\left(X_i^\ell \in \mathcal{X}_i^\ell | X_i^{i+d-1} = x_i^{i+d-1}\right)$ which is computable by recurrence through:*

$$B_i\left(x_i^{i+d-1}\right) = \sum_{x_{i+d}\in\mathcal{X}_{i+d}} \pi\left(x_i^{i+d-1}, x_{i+d}\right) B_{i+1}\left(x_{i+1}^{i+d}\right) \quad \forall i, 2 \leqslant i \leqslant \ell - d - 1 \tag{3}$$

with the initialization $B_{\ell-d}\left(x_{\ell-d}^{\ell-1}\right) = \sum_{x_\ell\in\mathcal{X}_\ell} \pi\left(x_{\ell-d}^{\ell-1}, x_\ell\right) \mathbb{I}_{\mathcal{X}_{\ell-d}^{\ell-1}}(x_{\ell-d}^{\ell-1})$. We then obtain that:

$$\mathbb{P}(X_1^\ell \in \mathcal{X}_1^\ell) = \sum_{x_1^d\in\mathcal{X}_1^d} \nu\left(x_1^d\right) B_1\left(x_1^d\right) \tag{4}$$

Proof. The proof is very similar to the one of Proposition 2.1 and is hence omitted. □

THEOREM 2.3 (MARGINAL DISTRIBUTIONS). *For all $x_1^\ell \in \mathcal{A}^\ell$ we have the following results:*

a) $\mathbb{P}\left(X_1^d = x_1^d, X_1^\ell \in \mathcal{X}_1^\ell\right) = \nu\left(x_1^d\right) B_1\left(x_1^d\right)$;
b) $\forall i, 1 \leqslant i \leqslant \ell-d-1, \mathbb{P}\left(X_i^{i+d} = x_i^{i+d}, X_1^\ell \in \mathcal{X}_1^\ell\right) = F_i\left(x_i^{i+d-1}\right) \pi\left(x_i^{i+d-1}, x_{i+d}\right) B_{i+1}\left(x_{i+1}^{i+d}\right)$;
c) $\mathbb{P}\left(X_{\ell-d}^\ell = x_{\ell-d}^\ell, X_1^\ell \in \mathcal{X}_1^\ell\right) = F_{\ell-d}\left(x_{\ell-d}^{\ell-1}\right) \pi\left(x_{\ell-d}^{\ell-1}, x_\ell\right)$;
d) $\forall i, 1 \leqslant i \leqslant \ell - d, \mathbb{P}\left(X_i^{i+d-1} = x_i^{i+d-1}, X_1^\ell \in \mathcal{X}_1^\ell\right) = F_i\left(x_i^{i+d-1}\right) B_i\left(x_i^{i+d-1}\right)$.

Proof. a), b), and c) are proved using the same conditioning mecanisms used in the proofs of propositions 2.1 and 2.2. One could note that Equation (4) is a direct consequence of a), while Equation (2) could be derived from c). Thanks to Equation (3), it is also clear that b) $\Rightarrow$ d) which achieves the proof. □

Like in the Hidden Markov Model (HMM) framework, one may use these marginal distributions to derive a Expectation-Maximization (EM) algorithm [3] allowing to compute the Maximum Likelihood Estimator (MLE) $\widehat{\theta} = \mathrm{argmax}_\theta \mathbb{P}_\theta\left(X_1^\ell \in \mathcal{X}_1^\ell\right)$ with $\theta \stackrel{\text{def}}{=} (\nu, \pi)$. To keep this article short, this point is left to the reader.

From now on we denote by $\mathbb{P}^C(A) \stackrel{\text{def}}{=} \mathbb{P}\left(A | X_1^\ell \in \mathcal{X}_1^\ell\right)$ the probability of an event A under the constraint that $X_1^\ell \in \mathcal{X}_1^\ell$.

THEOREM 2.4 (HETEROGENEOUS MARKOV CHAIN). *For all $x_1^\ell \in \mathcal{X}_1^\ell$ we have:*

$$\mathbb{P}^C\left(X_1^d = x_1^d\right) \propto \nu\left(x_1^d\right) B_1\left(x_1^d\right) \tag{5}$$

and $\forall i, 1 \leqslant i \leqslant \ell - d$

$$\mathbb{P}^C\left(X_{i+d} = x_{i+d} | X_i^{i+d-1} = x_i^{i+d-1}\right) \propto \pi\left(x_i^{i+d-1}, x_{i+d}\right) B_{i+1}\left(x_{i+1}^{i+d}\right). \tag{6}$$

This means that, under $\mathbb{P}^C$, X_1^ℓ *is a order* d *heterogeneous Markov chain which starting distribution is given by Equation (5) and transition matrix is given by Equation (6).*

Proof. Equation (5) is a direct consequence of Theorem 2.3a) and Equation (4). For Equation (6) we start by denoting $\mathbb{P}^C\left(X_{i+d} = x_{i+d} | X_i^{i+d-1} = x_i^{i+d-1}\right) = \mathbb{P}(A|B,C,D)$ with $A = \{X_{i+d} = x_{i+d}\}$, $B = \{X_i^{i+d-1} = x_i^{i+d-1}\}$, $C = \{X_1^i \in \mathcal{X}_1^i\}$, and $D = \{X_{i+1}^\ell \in \mathcal{X}_1^i\}$. Thanks to Bayes' formula we get that $\mathbb{P}(A|B,C,D) \propto \mathbb{P}(D|A,B,C) \times \mathbb{P}(A|B,C)$. We finally use the Markov property to get $\mathbb{P}(D|A,B,C) = B_{i+1}\left(x_{i+1}^{i+d}\right)$ and $\mathbb{P}(A|B,C) = \pi\left(x_i^{i+d-1}, x_{i+d}\right)$ which achieves the proof. □

3 Counting patterns

Let us consider here $\mathcal{W}$ a finite set of word over $\mathcal{A}$. We want to count the number N of positions where $\mathcal{W}$ occurs in our degenerated sequence. Unfortunatly, since the sequence itself is not observed, we study instead the number N of matching positions in the random sequence X_1^ℓ under $\mathbb{P}^C$. Thanks to Theorem 2.4 we hence need to establish the distribution of N over a heterogeneous order d Markov chain. To do so, we perform an optimal Markov chain embedding of the problem through a Deterministic Finite Automaton (DFA) as it is suggested in [4][5][6][7]. We use here the notations of [7]. Let $(\mathcal{A}, \mathcal{Q}, s, \mathcal{F}, \delta)$ be a *minimal* DFA that recognizes the language [4] $\mathcal{A}^*\mathcal{W}$ of all texts over $\mathcal{A}$ ending with an occurrence of $\mathcal{W}$. $\mathcal{Q}$ is a finite state space, $s \in \mathcal{Q}$ is the starting state, $\mathcal{F} \subset \mathcal{Q}$ is the subset of final states, and $\delta : \mathcal{Q} \times \mathcal{A} \to \mathcal{Q}$ is the transition function. We recursively extend the definition of δ over $\mathcal{Q} \times \mathcal{A}^*$ thanks to the relation $\delta(p, aw) \stackrel{\text{def}}{=} \delta(\delta(p,a), w)$ for all $p \in \mathcal{Q}, a \in \mathcal{A}, w \in \mathcal{A}^*$. We additionally suppose that this automaton is non d-ambiguous [5] which means that for all $q \in \mathcal{Q}$, $\delta^{-d}(p) \stackrel{\text{def}}{=} \left\{a_1^d \in \mathcal{A}_1^d, \exists p \in \mathcal{Q}, \delta\left(p, a_1^d\right) = q\right\}$ is either a singleton, or the empty set.

THEOREM 3.1 (MARKOV CHAIN EMBEDDING). *We consider the random sequence over* $\mathcal{Q}$ *defined by* $\widetilde{X}_0 \stackrel{def}{=} s$ *and* $\widetilde{X}_i \stackrel{def}{=} \delta(\widetilde{X}_{i-1}, X_i)\ \forall i, 1 \leqslant i \leqslant \ell$. *Under* $\mathbb{P}^C$, $(\widetilde{X}_i)_{i \geqslant d}$ *is a heterogeneous order 1 Markov chain over* $\mathcal{Q}' \stackrel{def}{=} \delta(s, \mathcal{A}^d\mathcal{A}^*)$ *such as, for all* $p, q \in \mathcal{Q}'$ *and* $1 \leqslant i \leqslant \ell - d$ *the starting distribution* $\mu_d(p) \stackrel{def}{=} \mathbb{P}^C\left(\widetilde{X}_d = p\right)$ *and transition matrix* $T_{i+d}(p,q) \stackrel{def}{=} \mathbb{P}^C\left(\widetilde{X}_{i+d} = q | \widetilde{X}_{i+d-1} = p\right)$ *are given by:*

$$\mu_d(p) = \begin{cases} \mathbb{P}^C\left(X_1^d = a_1^d\right) \text{ if } \exists a_1^d \in \mathcal{A}^d, \delta\left(s, a_1^d\right) = p \\ 0 \qquad\qquad \text{else} \end{cases}; \tag{7}$$

$$T_{i+d}(p,q) = \begin{cases} \mathbb{P}^C\left(X_{i+d} = b | X_i^{i+d-1} = \delta^{-d}(p)\right) \text{ if } \exists b \in \mathcal{A}, \delta(p,b) = q \\ 0 \qquad\qquad\qquad\qquad\qquad \text{else} \end{cases}. \tag{8}$$

Proof. This comes from a direct application of Theorem 2.4 as well as results from [6] or [7]. □

COROLLARY 3.2 (MOMENT GENERATING FUNCTION). *The moment generating function* $F(y)$ *of the random number* N *over* $\mathbb{P}^C$ *is given by:*

$$F(y) \stackrel{def}{=} \sum_{k=0}^{+\infty} \mathbb{P}^C(N = k)\, y^k = \mu_d \left[\prod_{i=1}^{\ell-d} (P_{i+d} + yQ_{i+d})\right] \mathbf{1} \tag{9}$$

[4] $\mathcal{A}^*$ denotes the set of all (possibly empty) texts over $\mathcal{A}$.

[5] a DFA having this property is also called a d-th order DFA in [6].

where $\mathbf{1}$ *is a column vector of ones and where, for all* $1 \leqslant i \leqslant \ell - d$, $T_{i+d} = P_{i+d} + Q_{i+d}$ *with* $P_{i+d}(p,q) \stackrel{def}{=} \mathbb{I}_{q \notin \mathcal{F}} T_{i+d}(p,q)$ *and* $Q_{i+d}(p,q) \stackrel{def}{=} \mathbb{I}_{q \in \mathcal{F}} T_{i+d}(p,q)$ *for all* $p, q \in \mathcal{Q}'$.

Proof. Since Q_{i+d} contains all counting transitions, we keep track of the number of occurrences by associating a dummy variable y to these transitions. We hence just have to compute the marginal distribution at the end of the sequence and sum up the contribution of each state. See [4][5][6][7] for more details. □

4 Example

Let us consider the dataset `est_pro_01` which is described in Table 1. Here is the transition matrix over of a order $d = 1$ homogeneous Markov model over $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ estimated on this dataset using MLE (though the EM algorithm):

$$\widehat{\pi} = \begin{pmatrix} 0.3337 & 0.1706 & 0.2363 & 0.2595 \\ 0.2636 & 0.2609 & 0.1775 & 0.2980 \\ 0.2946 & 0.2218 & 0.2666 & 0.2169 \\ 0.2280 & 0.2413 & 0.2106 & 0.3201 \end{pmatrix}.$$

Since only 1% of the dataset is degenerated, we observe little difference between this rigorous estimate and one obtain though a rough heuristic (like discarding all degenerated positions in the data).

However, this result should not be taken as a rule, especially when considering more degenerated sequences (*e. g.* with 10% degenerated positions) and/or higher order Markov models (*e. g.* $d = 4$).

pattern	naive count	lower bound	5%-percentile	median	95%-percentile	upper bound
`GCTA`	715	715	727	733	740	824
`TTAGT`	197	197	201	205	209	253
`TTNGT`	839	853	874	881	889	1005
`TRNANNNSTM`	472	477	488	493	498	535

Table 2. Distribution of patterns in the degenerated IUPAC sequences from `est_pro_01`. The "naive count" of pattern occurrences is the one obtained by discarding all degenerated positions in the dataset. Since the observed distribution is discrete, percentiles and median are rounded to the closest value.

Using this model, it is possible to study the *observed distribution* of a pattern in the dataset by computing though Corollary 3.2 the distribution of its random number of occurrence N under the constrained probability $\mathbb{P}^C$. Table 2 compares a "naive" number of occurrences (obtained by discarding all degenerated positions in the data) to the observed distribution. Despite the fact that only 1% of the data are degenerated, we can see that there is a great differences between our naive approach and the real observed distribution. For example, if we consider the simple pattern `GCTA` we can see that the naive count of 715 occurrences lies well outside the 90% credibility interval $[727, 740]$. And we have similar results for the other considered patterns. For more complex patterns like `TTNGT` the difference between the naive count and the observed distribution is even more dramatic since 839 does not even belong to the support $[853, 1005]$ of the observed distribution. This is due to the fact that the *string* `TTNGT` actually occurs $853 - 839 = 14$ times in the dataset. Since our naive approach discard all positions in the data where a symbol other than `A`, `C`, `G` or `T` appears, these

14 occurrences are hence omitted. Finally, let us point out that thanks to the optimal Markov chain embedding provided by the DFA-based approach presented above, we are here able to deal with relatively complex patterns like TRNANNNSTM. One might suggests other *ad hoc* counting heuristics than the "naive" one introduced above. For example, one can preprocess the dataset by replacing all degenerated symbols by the most frequent letter in the corresponding subset. In our example, such an approach leads to the following countings: 732 for GCTA, 211 for TTAGT, 853 for TTNGT, and 505 for TRNANNNSTM. If this heuristic gives an interesting result for the first pattern (counting close to the median), it is unfortunately not the case for the other ones. Moreover, it is difficult to predict the bias introduced by this particular heuristic since it can either lead to under- or over-coutings depending on the pattern.

5 Conclusion

In this paper, we provide a rigorous way to deal with the distribution of Markov chains over a finite alphabet $\mathcal{A}$ under the constraint that each position X_i of the sequence belongs to restricted subset $\mathcal{X}_i \subset \mathcal{A}$. We provide a Forward-Backward framework to compute marginal distributions and derive from it a EM estimation procedure. We also prove that the resulting constrained distribution is a heterogeneous Markov chains and provide explicit formulas to recursively compute its transition matrix. Thanks to this result, it is possible to apply known DFA-based methods from pattern theory to study the distribution of a pattern of interest in this constrained sequence, hence providing a trustful observed distribution for the pattern number of occurrences. This information may then be used to derive a p-value p for a pattern by combining p_n the p-value of the observation of n occurrences in a unconstrained dataset with the observed distribution through formulas like $p = \sum_n p_n \mathbb{P}^C(N = n)$. One should note that the approach we introduce here may have more application than just counting patterns in IUPAC sequences. For example, one might use a similar approach to take into account the occurrences positions of known patterns of interest thus allowing to derive distribution of patterns conditionnally to a possibly complex set of other patterns. One should also point out that the constraint $X_i \in \mathcal{X}_i$ should easily be complexified, for example by considering a specific distribution over $\mathcal{X}_i$. For instance, such a distribution may come from the posterior decoding probabilities of a sequencing machine.From the computational point of view, it is essential to understand that the heterogeneous nature of the Markov chain we consider forbid to use classical computational tricks like power computations. The resulting complexity is hence linear with the sequence length ℓ rather that logarithmic. However, one should expect a dramatic improvement of the method by restricting the use of heterogeneous Markov models only in the vicinity of degenerated positions.

References

[1] IUPAC, International Union of Pure and Applied Chemistry, http://www.iupac.org, 2009.
[2] EMBL Nucleotide Sequence Database, http://www.ebi.ac.uk/embl/, 2009.
[3] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Stat. Society. Series B*,39(1):1-38, 1977.
[4] P. Nicodème, B. Salvy and P. Flajolet, Motif statistics. *Theoretical Com. Sci.*, 287(2):593-617, 2002.
[5] M. Chrochemore and V. Stefanov, Waiting time and complexity for matching patterns with automata. *Info. Proc. Letters*, 87(3):119-125, 2003.
[6] M. E. Lladser, Mininal Markov chain embeddings of pattern problems. *Information Theory and Applications Workshop, 2007*, 251-255, 2007.
[7] G. Nuel. Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. *J. of Applied Prob.*, 45(1):226-243, 2008.

How to measure the robustness of bacterial genome comparisons?

Hugo Devillers[1], Hélène Chiapello[1], Meriem El Karoui[2], Sophie Schbath[1]

[1] Unité Mathématique, Informatique & Génome, UR1077 INRA,
Domaine de Vilvert, F-78350, Jouy-en-Josas France
[2] Unité Bactéries Lactiques et Pathogènes Opportunistes, UR888 INRA,
Domaine de Vilvert, F-78350, Jouy-en-Josas France
{hugo.devillers, helene.chiapello, meriem.el_karoui,
sophie.schbath}@jouy.inra.fr

Abstract: *The number of studies dealing with complete bacterial genome comparisons steadily increases. They allow us to gain insight into the molecular mechanisms involved in the evolution of bacterial genomes such as DNA exchanges. There exist several software tools and methods to align complete genomes and to determine conserved and variable regions. However, statistical methods to evaluate these tools are lacking. To fill this gap, two local scores for measuring the robustness of the comparisons of bacterial genomes are proposed. The calculation procedures of these scores are first presented and their interest is then discussed from two illustrative examples.*

Keywords: Comparative genomics, bacteria, conserved/variable segments, robustness, complete genome.

1 Introduction

The number of complete bacterial genome sequences available in public databases has considerably increased since the publication, in 1995, of the genome of *Haemophilus influenzae* that was the first bacterium to be completely sequenced [1]. There are currently more than 700 bacterial genomes entirely sequenced, representing about 250 distinct genera, and more than 1,200 other genomes will be available soon (see: http://www.ncbi.nlm.nih.gov/Genomes/, December 2008). Comparison of these genomes allows us to address new questions about their structure and their evolution [2]. Moreover, since the publication of a second strain of *Helicobacter pylori* in 1999 [3], the availability of genomes of closely related bacterial strains has rapidly increased. This offers new opportunities to gain insight into the understanding of short-term evolutionary processes, especially at the molecular level.

A comparison of two closely related bacterial genome sequences was performed by Hayashi *et al.* in 2001 [4]. An alignment of the two complete genomes of the enterohemorrhagic *Escherichia coli* O157:H7 Sakai strain and the *E. coli* K-12 MG1655 laboratory strain was performed. It allowed the determination of a highly conserved sequence between the two genomes, called the conserved backbone of the *E. coli* chromosome, which was interrupted by several DNA segments that were variable from one strain to the other. The backbone/variable segment structure is named segmentation. Its analysis is of great interest to study the molecular mechanisms involved in the

dynamics of bacterial genome evolution. Thus, for example, segments from the conserved backbone, which may correspond in large part to the common ancestral strain, have been shown to be enriched in functional DNA motifs [5]. Variable segments that may be associated to strain-specificities, are particularly relevant to study horizontal transfers, as they are probably associated to mobile elements such as prophages [6]. Consequently, the segmentation (backbone/variable segments) must be accurately determined. There exist various software tools to compare and to align bacterial genomes [2] and several databases store pre-computed comparisons such as xBASE [7] and MOSAIC [8].

The success of sequence alignment methods, such as BLAST or FASTA, lies, in part, in the evaluation of the statistical significance of the alignment score they provide. The genome comparison tools cited above generally suffer from a lack of statistical methods to evaluate their results [9]. To fill this gap, we propose two local scores measuring the robustness of the segmentations of bacterial genomes. In this paper, the calculation procedures of these two scores are first presented and their interest is then stressed from two illustrative examples.

2 Measuring the Segmentation Robustness

Here we present a method to measure the robustness of a segmentation (*i.e.*, a backbone/variable segment structure) obtained from the comparison of two genomes. Our method is based on a simulation process that aims at randomly perturb the original genomes.

2.1 Simulation Process

The determination of bacterial genome segmentation is generally based on the detection of the common elements between the compared sequences. Thus, to measure the robustness of such a procedure, it is relevant to perturb only conserved regions rather than random sequences chosen from the whole genomes. We therefore focus on maximal exact matches (MEMs), which correspond to common sequences between the compared genomes that cannot be extended (whose length is maximal). It is noteworthy that MEMs are frequently used as anchors to align complete genomes [10]. The nucleotides corresponding to a user defined proportion of these MEMs are randomly perturbed. Three types of perturbations are defined: 1) Deletions, MEM's positions are simply deleted; 2) Inversions, a MEM sequence is reverse-complemented and reinserted at the same position; 3) Double translocations, two MEM sequences are switched. Perturbations are applied separately in each compared genome, so that the process is symmetrical. The segmentation of the perturbed genomes is then computed and stored in a database. The process is repeated a sufficient number of times to ensure the statistical reliability of the scores defined below.

2.2 Score Definition

The measurement of robustness is based on the evaluation of the differences between the original segmentation (*i.e.*, the backbone/variable segment structure) and the segmentations computed with the perturbed genomes. Two scores are derived, one focusing on the nucleotide robustness, the other one on the robustness of the segments. Considering the nucleotide *i* from one of the original genomes (either from a backbone or a variable segment of the original segmentation), the nucleotide score is defined as follows:

$$S_{nuc}(i) = \frac{\#\{simulations \mid i \in variable\ segment\}}{\#\{total\ simulations\}}.$$

It is equal to the proportion of simulations in which the nucleotide *i* is assigned in a variable segment. Thus, S_{nuc} varies between 0 and 1. Its interpretation is the following: if $S_{nuc}(i)$ is near 1 then *i* is likely to belong to a variable segment.

Considering the segment *g* of the original segmentation (*i.e.*, the non-perturbed segmentation), the segment score is defined by:

$$S_{seg}(g) = \frac{1}{|g|} \sum_{i \in g} S_{nuc}(i),$$

where |*g*| denotes the number of nucleotides in segment *g*. It is equal to the average of the nucleotide scores of the nucleotides belonging to segment *g*. Thus, if $S_{seg}(g)$ is close to 1 then the segment g is likely to be a robust variable segment.

3 Application to Two Segmentations in the *Escherichia coli* Species

3.1 Dataset Selection

We first compared the *E. coli* enterohemorrhagic O157:H7 Sakai strain and the *E. coli* K-12 MG1655 laboratory strain. The corresponding segmentation is available in the MOSAIC database (http://genome.jouy.inra.fr/mosaic/). This choice relies on the fact that this segmentation has been intensively studied and compares well to a manually curated dataset [4]. We also used a second segmentation based on the comparison of two *E. coli* K-12 strains: K-12 MG1655 and K-12 W3110. The segmentation was performed using the strategy developed for the MOSAIC database. Because these two genomes are almost identical, this segmentation is expected to be roughly constituted by a unique backbone segment. Surprisingly, it is not the case as 40% of the genomes appear in variable segments. This suggests that the segmentation strategy might need to be modified for such closely related genomes (see below). These two *E. coli* segmentations were used here to illustrate the interest of the two scores.

3.2 Nucleotide Score

For each selected segmentation, S_{nuc} was computed. After a preliminary investigation, it was decided to perturb 33% of the MEMs using a combination of the three types of perturbations described in section 2.1 and to perform 100 simulations. S_{nuc} values were then plotted for all the nucleotides of each genome. Three examples representative of the different score profiles are shown in Fig. 1.

Fig. 1A shows a first example for the K-12/Sakai strain segmentation, which is focused on a 5,000 bp variable segment. Along this region, S_{nuc} is equal to 1 for the variable segment and sharply decreases at the surrounding backbone segments. This strongly suggests that the nucleotides of the focused segment really belong to a variable segment. Fig. 1B displays another variable segment from the K-12/Sakai strain segmentation. Values of S_{nuc} indicate that although the assignment of the nucleotides to this variable segment is globally robust, the assignment for those located at the boundaries of the segment is less robust than for the others.

Fig. 1C depicts S_{nuc} results for a variable segment of the comparison between the two *E. coli* K-12 strains. The very low S_{nuc} values along this segment reveal that the later is not robust and lead to suppose that it cannot be considered as a variable segment.

These three above examples of S_{nuc} profiles indicate that the nucleotide score allows us to precisely analyze the robustness of a segmentation along each nucleotide of a genome. It facilitates

the detection of non or partially robust segments. Similar analyses were done along backbone segments (not shown) and indicate that this score is also useful to analyze backbone segments.

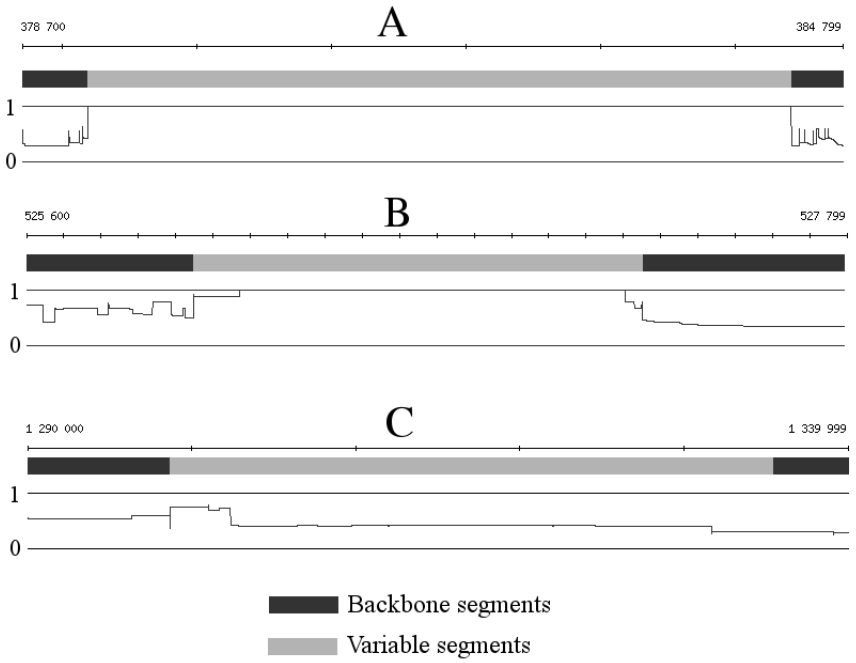


Figure 1. Nucleotide scores for three variable segments along the *E. coli* K-12 MG1655 genome from the segmentation of K-12/Sakai strains (A and B) and from the segmentation of the two K-12 strains (C). The axis at the top gives the nucleotide positions, the black and gray line shows the computed segmentation, and the curve (varying from 0 to 1) displays the nucleotide scores.

3.3 Segment Score

Computation of the segment scores (S_{seg}) was also performed on the two selected segmentations of the *E. coli* species. Fig. 2A displays the histogram of S_{seg} values for all the segments of K-12 MG1655 from the comparison of K-12/Sakai strains. This segmentation contains 617 variable segments and 618 backbone segments. The score distribution presents two peaks, one for the variable segments and the other for the backbone segments. Most of the variable segments (in gray in Fig. 2A) have a score between 0.99 and 1, indicating that they are robust. The backbone segments (in black in Fig. 2A) most often have a score ranging between 0.3 and 0.4. They are also probably robust. Indeed, the backbone being mainly constituted of MEMs, their percentage of perturbation will determine the expected value of a robust score for a backbone segment. Because in this study 33% of the MEMs were perturbed, robust backbone segment scores are expected to be around 0.33. Thus, from a rapid inspection of Fig. 2A, we can easily conclude that the whole segmentation of K-12/Sakai strains is robust.

Conversely, it is not the case for the segmentation of the two substrains of *E. coli* K-12 strains (Fig. 2B). This figure clearly shows that for most of the variable and backbone segments, the score values correspond to a low robustness. This is in agreement with the fact that the predicted segmentation contains unexpected variable segments while a unique backbone segment was expected. As a result we can conclude that the whole segmentation of the two substrains of *E. coli* K-12 strains is not robust.

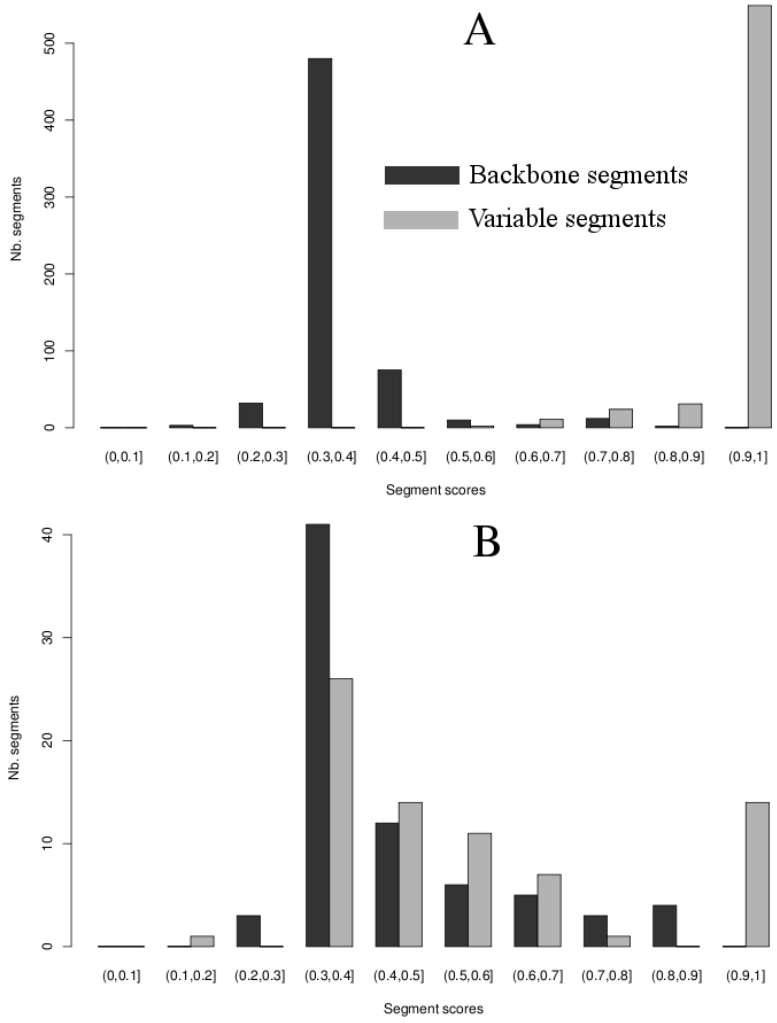


Figure 2. Segment scores for the segmentations of the *E. coli* K-12 MG1655 from the comparison of K-12/Sakai strains (A) and from the comparison of the two K-12 strains (B).

4 Concluding Remarks

To our knowledge, this study is the first attempt to statistically determine the robustness of bacterial genome segmentations. The two proposed scores, routed on classical statistics are simple to compute and easy to interpret. The examples presented here show that the proposed scores are able to distinguish robust and non robust segmentations. A statistical test will then be designed for this purpose. The nucleotide score (S_{nuc}) also allows to detect short non robust regions among a generally robust segmentation.

Such encouraging results have been also obtained from the analysis of several other segmentations from the MOSAIC database (data not shown). This suggests that the scores developed here could be used at a larger scale, for example on all comparisons stored in the MOSAIC database. To further validate our approach, we are also performing simulation studies on artificial genomes for which the segmentation is known.

Comparison of multiple strains of a single species has also yielded the concept of species pan-

genome as a measure of the whole gene repertoire that can pertain to a given bacterium [11]. Briefly, genes of the pan-genome are divided into three categories. The core-genome groups genes shared by all the strains, the dispensable genes correspond to those that are not present in each strain and last, the specific genes are observed in only one strain. In this context, it should be interesting to see whether genes of the core-genome belong to robust backbone segments as determined by the score calculations. This will be investigated in future works.

Acknowledgements

We are grateful to the INRA MIGALE platform (http://migale.jouy.inra.fr) for providing computational resources. We thank Annie Gendrault for her valuable help in database management. This work was supported by the French ANR (Agence Nationale de la Recherche) project CoCoGen (BLAN07-1_185484).

References

[1] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, K.S. McKenney, G. Sutton, W. Fitzhugh, C. Fields, J.D. Gocyne, J. Scott, R. Shirley, L. Liu, A. Glodek, J.M. Kelley, J.F. Weidman, C.A. Phillips, T. Spriggs, E. Hedblom, M.D. Cotton, T.R. Utterback, M.C. Hanna, D.T. Nguyen, D.M. Saudek, R.C. Brandon, L.D. Fine, J.L. Fritchman, J.L. Fuhrmann, N.S.M. Geoghagen, C.L. Gnehm, L.A. McDonald, K.V. Small, C.M. Fraser, H.O. Smith and J.C. Venter, Whole-genome random sequencing and assembly of *Haemophilus influenzae*. Science, 269:496-512, 1995.

[2] D. Field, G. Wilson and C. van der Gast, How do we compare hundreds of bacterial genomes? Curr. Opin. Microbiol., 9:499-504, 2006.

[3] R.A. Alm, L.S. Ling, D.T. Moir, B.L. King, E.D. Brown, P.C. Doig, D.R. Smith, B. Noonan, B.C. Guild, B.L. deJonge, G. Carmel, P.J. Tummino, A. Caruso, M. Uria-Nickelsen, D.M. Mills, C. Ives, R. Gibson, D. Merberg, S.D. Mills, Q. Jiang, D.E. Taylor, G.F. Vovis, T.J. Trust, Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature, 397:176-180, 1999.

[4] T. Hayashi, K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C.G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasakawa, N. Ogasawara, T. Yasunaga, S. Kuhara, T. Shiba, M. Hattori and H. Shinagawa, Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res., 8:11-22, 2001.

[5] D. Halpern, H. Chiapello, S. Schbath, S. Robin, C. Hennequet-Antier, A. Gruss and M. El Karoui, Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling. PLoS genet., 9:153-160, 2007.

[6] H. Chiapello, I. Bourgait, F. Sourivong, G. Heuclin, A. Gendrault-Jacquemard, M.A. Petit and M. El Karoui, Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. BMC Bioinformatics, 6:171-180, 2005.

[7] R.R. Chaudhuri and M.J. Pallen, xBASE, a collection of online databases for bacterial comparative genomics. Nucleic Acids Res., 34:335-337, 2006.

[8] H. Chiapello, A. Gendrault, C. Caron, J. Blum, M.A. Petit and M. El Karoui, MOSAIC: an online database dedicated to the comparative genomics of bacterial strains at the intra-species level. BMC Bioinformatics, 9:498-506, 2008.

[9] W. Miller, Comparison of genomic DNA sequences: solved and unsolved problems. Bioinformatics, 17:391-397, 2001.

[10] M. Höhl, S. Kurtz and E. Ohlebusch, Efficient multiple genome alignment. Bioinformatics, 18:S312-S320, 2002.

[11] A. Muzzi, V. Massignani and R. Rappuoli, The pan-genome: towards a knowlegde-based discovery of novel targets for vaccines and antibacterials. Drug Discov. Today, 12:429:439, 2007.

Improved sensitivity and reliability of anchor based genome alignment

Raluca Uricaru[1], Célia Michotey[2], Laurent Noé[3], Hélene Chiapello[2], Eric Rivals[1]

[1] LIRMM, CNRS and Université de Montpellier 2
161, rue Ada, 34392 Montpellier cedex 5, France
{uricaru, rivals}@lirmm.fr
[2] INRA UR1077, Unité Mathématique, Informatique & Génome,
Domaine de Vilvert, 78352, Jouy-en-Josas, France
{helene.chiapello, celia.michotey}@jouy.inra.fr
[3] LIFL - INRIA Université de Lille I, Villeneuve d'Ascq, France
noe@lifl.fr

Abstract: *Whole genome alignment is a challenging problem in computational comparative genomics. It is essential for the functional annotation of genomes, the understanding of their evolution, and for phylogenomics. Many global alignment programs are heuristic variations on the anchor based strategy, which relies on the initial detection of similarities and their selection in an ordered chain. Considering that alignment tools fail to align some pairs of bacterial strains, we investigate whether this is intrinsically due to the strategy or to a lack of sensitivity of the similarity detection method. For this, we implement and compare* 6 *programs based on three different detection methods (from exact matches to local alignments) on a large benchmark set. Our results suggest that the sensitivity of well known methods, like MGA or Mauve, can be greatly improved in the case of divergent genomes if one exploits spaced seeds at the detection phase. In other cases, such methods yield alignments that cover nearly the whole genome. Then, we focus on global reliability of alignments: should an aligned pair of segments be included in the global genome alignment? We investigate this reliability according to both the segment "alignability" and to inclusion of orthologs. Again, we provide evidence that for both close and divergent genomes, one of our programs, YH, achieves alignments with sometimes a lower coverage, but a higher inclusion of orthologs. It opens the way to the first reliable alignments for some highly divergent species like Buchnera aphidicola or Prochlorococcus marinus.*

Keywords: Global genome alignment, anchor based strategy, spaced seeds

1 Introduction

Whole genome comparisons offer a unique opportunity to investigate globally the mechanisms of evolution in closely related species, are a key to the inference of functional elements in both coding and non coding regions, and serve as a basis in phylogenomics [1,2]. In particular, the conserved parts of genomes, forming the so called *backbone segments*, indicate the biological components common to several species or strains, while differences in sequences, *variable segments*, are likely responsible for what distinguishes them (*e.g.*, pathogenic islands). Genome alignment can deliver both at once.

Due to the genome sizes and to the task complexity, genome alignment tools implement heuristic algorithms. The most used scheme is the **anchor based strategy** (*e.g.*, [3,4,5]), which operates in four phases. It starts by detecting an initial set of pairwise similarity regions (phase 1) and, through a *chaining* phase, selects a non-overlapping maximum-weighted subset of those similarities (phase 2), called *anchors*. Phases 1 and 2 are recursively applied to each pair of yet unaligned regions (phase 3). The last phase consists in systematically applying classical heuristic alignment tools (*e.g.*, ClustalW) to all short region pairs still left unaligned.

The Mosaic database stores the alignments of backbone segments for every pair of strains of the same bacterial species. The backbones are obtained by first aligning the genomes with either MGA or Mauve (two anchor based tools), then by post-processing the alignment to remove segment pairs whose percentage of identity falls below 76%. This post-processing, although based on an arbitrary threshold, is still applied to avoid unreliable alignments. Moreover, some pairs of strains are absent from the database because the backbones covered less than 50% of the genome. It is unanswered whether cases of small coverage are due to a lack of sensitivity of the methods or to an intrinsic limitation of the strategy.

Here, we investigate this issue by implementing and comparing six methods that combine three similarity detection methods and two chaining algorithms, and by comparing the results on a large benchmark made of all pairwise intra-species bacterial genomes. As they simulate the first 2 phases of the strategy, those methods can also be compared to MGA and Mauve results. It turns out that one of the programs that exploits spaced seeds to search for similarity regions, YH, allows to align divergent collinear genomes, for which MGA and Mauve failed to produce reliable alignments. Moreover, when comparing the proportion of orthologous genes included in the alignments, YH seems to overcome some reliability problems encountered by other methods, including on well-known cases like *E. coli*.

In the sequel, Section 2 presents our programs, the benchmark data, and establishes a protocol for the evaluation of global alignment. In Section 3, we evaluate the performance of those methods from both computational and biological view-points, while we discuss the results in Section 4.

2 Methods

Genome Alignment Programs MGA and Mauve are two archetypal anchor based alignment tools, are widely used, documented and proved to be more accurate than MUMer and SLagan [5,6]. They differ by two aspects: in phase 1, MGA searches for similarity regions that are *maximal exact matches* (MEMs) with the program Vmatch [3], while Mauve finds *approximate matches* using a special type of spaced seeds [4]. In phase 2, MGA executes Chainer [7], a program that selects the highest scoring non-overlapping set of collinear matches (*consistent chain* [5]); Mauve uses a greedy breakpoint elimination algorithm [8] that generates an approximate solution to the maximum-weighted non collinear anchoring problem. Hence, MGA treats collinear genome pairs, while Mauve handles rearrangements.

Our 6 programs combine one of 3 similarity detection methods (Vmatch, Blast v2, Yass) and one of 2 chaining algorithms (Chainer and Hierarchical chainer). Contrarily to Vmatch, Blast v2 and Yass find similarities that are local alignments with either contiguous or spaced seeds [9]. We named our 2×3 combinations by the initials of the methods they combine: VC, BC and YC with Chainer, VH, BH and YH with Hierarchical Chainer. By comparing those, we can measure the impact of each element on the final alignment.

The *Hierarchical chainer* implements a greedy chaining that allows for limited overlaps, in place inversions, and privileges region pairs with stronger similarities [10]. Similarities are ordered by decreasing *numbers of identities* (nid) and processed in groups according to several intervals of nid, starting with the largest ones. In each group, we consider first similarities located on the dotplot main diagonal (*i.e.*, shift of 0) and continue with increasing shifts. collinear similarities are being chained from the left end on the reference sequence, in a greedy manner.

Genome Sequences and Comparisons We considered all (236) pairs of bacterial strains of the same species whose complete genomes are available in GenomeReviews database as of mid-2008 [11]. The Mauve and MGA alignments, and the backbone segments positions for each pair were obtained from the Mosaic database [12], except for 37 pairs corresponding to five divergent species (*Buchnera aphidicola, Prochlorococcus marinus, Pseudomonas fluorescens, Rhodopseudomonas palustris* and *Synechococcus sp*) that were recomputed with the same protocol, since excluded from the database due to poor backbone coverage. For all pairs, we compute the alignments with our 6 programs. For 13 pairs, Vmatch yields erroneous results (detects non existing MEMs), which were excluded from further analysis. The backbone according to YH is the intersection of the set of anchors on each genome.

Criteria for Evaluation To compare alignments, all genome aligner publications use global quantitative criteria like the *percentage of identity* (%id) and the *coverage*, however not necessarily with the same definition. The usual definition of the %id, percentage of identical base pairs over the total alignment length (as in Mosaic), makes it incomparable between alignments. We define the *coverage* as the total length of aligned segments, and the %id as the ratio of identical bases in aligned segments (of the coverage) *over the genome length*.

To measure the reliability of the genome alignments, we compare their intersection with the sets of orthologous genes as defined in the OMA database [13]. We retrieved from OMA the list of orthologous genes and their positions for 12 pairs. For each, we compute the number and % of the genomic sequence of orthologous genes included in the backbone.

3 Results

The six programs were applied on every pair of intra-species bacterial genomes (see Section 2). The results were compared to those obtained using MGA, Mauve (collected from Mosaic) with respect to the criteria defined above. Result tables and additional information can be found at the following location: http://www.lirmm.fr/~uricaru/Appendix_JOBIM09.html (Appendix).

Present Achievements
The first striking result lies in the difference of coverage between different species obtained by MGA and Mauve. For some species all pairwise alignments cover more than 90% of the genome (*e.g.*, on *Streptococcus thermophilus*), while in others the coverage is below 10% (*e.g.*, *Synechococcus sp*). One also observes, but more rarely, species for which the coverage of both methods varies greatly among pairs of strains (for *P. marinus*, the coverage of MGA varies in $[0, 78]\%$ and that of Mauve in $[6, 96]\%$).

It is clear that these programs succeed in aligning some genome pairs and fail in others, which could be due either to a high level of divergence that makes the sequences unalignable (see Appendix) or to a methodological failure in detecting similarity regions or in chaining.

Local Similarities (LS) vs MEMs

The similarity detection phase is mainly responsible for the sensitivity of an anchor based method. Indeed, the chaining phase only discards potential anchors, however it may be unadapted to the type of similarities used. Here, to assess the impact on sensitivity, we compare the genome coverage obtained after the phases 1 and 2 of three different similarity detection methods combined with two chaining algorithms on a large panel of genome comparisons. Similarities are either short exact matches (MEMs), BLAST local alignments or local alignments based on spaced seeds (YASS). Figure 1 shows the difference of genome coverage between pairs of methods as box plots.

The left plot, which compares the effect of MEMs versus spaced seeds LS combined with Chainer, demonstrates that a classical chaining algorithm is unadapted to LS. This is due to overlaps between long local similarities, which are prohibited in the chain and cause Chainer to discard large alignment regions, especially for highly similar genomes. We thus designed a new chaining method allowing for overlaps, called Hierarchical chaining (see Section 2).

The central part of Figure 1 shows which chaining method suits a given type of similarities. Chainer does well with MEMs (VH-VC), Hierarchical performs in average better than Chainer with BLAST LS (BH-BC), and always surpasses it with spaced seeds LS (YH-YC). The right part compares the combination YH with the other best combinations. It clearly shows that YH surpasses all other methods in coverage and can even achieve important differences. Let us take the case of *P. marinus* strains $CP000111\_GR$ vs $CP000095\_GR$ as a running example of divergent strains: YH obtains 57% coverage, while BH covers 3% of the genome and VC not even 1%.

Our comparisons provide clear evidence that using spaced seeds in the similarity detection phase improves the coverage, and therefore the global sensitivity of the anchor strategy. Thus, in some cases, alignment failure was due, indeed, to a lack of sensitivity of the anchor detection phase.

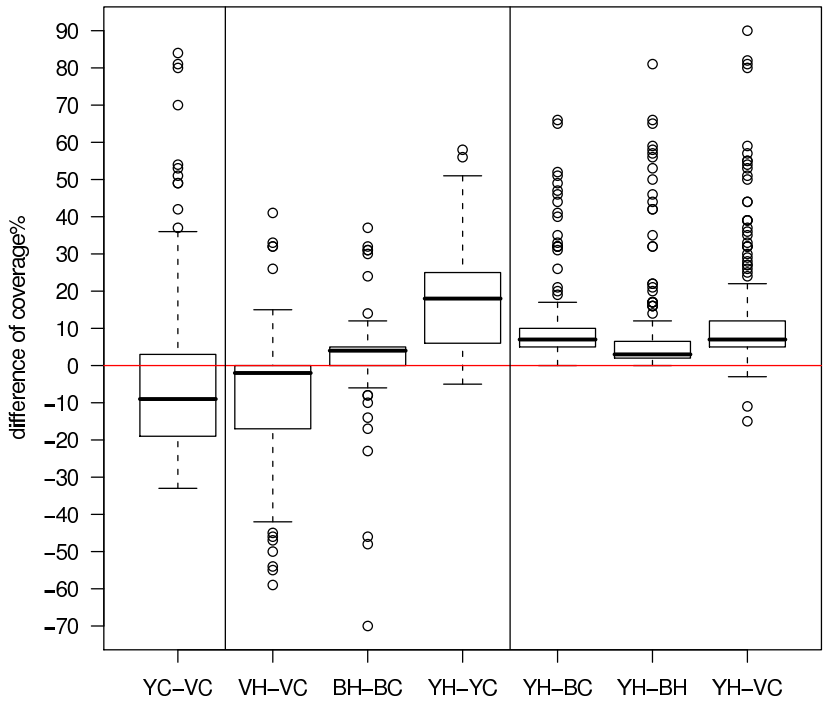


Figure 1: Boxplots of genome coverage differences between methods over 236 genome pairs. *E.g.*, YH−YC means, for each pair, the genome coverage of YH method minus that of YC in %. In a boxplot, circles are outliers. Left part: YC−VC coverage. Central part: comparison of each similarity detection method combined either with Chainer or our Hierarchical chaining. The 4th boxplot plus the right part: YH compared to other combinations; YH obtains larger coverage in the vast majority of cases over all four other combinations.

YH vs MGA/Mauve
Although MGA and Mauve execute two additional steps compared to our programs, the comparison of their results allows to see which proportion of the genome can be aligned solely with the chain of anchors and how it contributes to the percentage of identities.

On the 236 genome pairs, the coverage difference YH-MGA varies from -17% to 99%, with an average of 7.2%, and is zero or positive for 140 pairs and $< -2\%$ for only 10 pairs. The difference in %id varies from -16% to 99% with an average of 7%, and is zero or positive for 183 pairs and $< -2\%$ for only 2 pairs. Hence, the hierarchical program either achieves a result similar to MGA or improves on it in both aspects: coverage and %id.

In average, over the 236 genome pairs, the alignments of Mauve cover 13% more nucleotides than that of YH (variation within $[-14, 69]\%$) and have nearly 10% more identities. This is mostly due to its ability to handle rearrangements. However, as detected in Mosaic, a high coverage sometimes hides unreliable alignments. The segments that are aligned with ClustalW in the fourth phase (these are termed *aligned gaps* by Mauve) do not necessarily share sequence similarity and are often unreliably aligned. We investigated whether unreliable segments have a high impact on the coverage especially for highly divergent strains.

Our running example with a pair of *P. marinus* strains is in fact a typical situation for the divergent bacterial cases. In this case, Mauve covers in average with 27% more than YH (84%, corresponding to 1491kb vs 57% corresponding to 1012kb), while the difference in identity percentage is only of 8% in its favour. As both the coverage and the %id are ratios over the genome length, we can say that it covers 27 additional % (479kb) of the genome with only 8% more identities (142kb). It suggests that some pairs of aligned segments could well be false positives (*i.e.*, should not be part of the alignment). Indeed, by plotting the cumulative coverage with segments below a given threshold of %id, we found that Mauve covers 22, resp. 30%, with segments whose %id lies ≤ 50, resp. $\leq 55\%$.

Finally, we looked at the reliability and the accuracy of our backbones with a biological view-point. For this we compared the percentage of nucleotides from orthologous genes and the number of such genes included in YH, MGA, or Mauve backbones. In our *P. marinus* example, 60% of the nucleotides are part of YH backbone, compared to only 7% for Mauve and 3% for MGA. Even for the well studied *E. coli* comparison (K12 vs Sakai), where all three tools report a coverage 80% on Sakai genome, YH completely includes in its backbone 8% more orthologous genes than the other tools do. Even if it can be improved, this suggests that YH gives accurate and reliable backbones, with more precise segment bounds.

4 Discussion

In this work, we conducted one of the first evaluations of anchor-based genome alignment methods on a large set of intra-species bacterial genome alignments. For this, we propose a protocol and implement several programs performing only the first two steps of the anchor based strategy.

First, it appears that even for short, closely related, and sometimes collinear genomes, pairwise alignment is incompletely solved by nowadays programs. Second, the anchor chain they compute can be improved by using local alignments instead of shorter exact or approximate matches as similarities, provided that the chaining algorithm authorises overlaps between adjacent anchors. This improvement measured in terms of genome coverage and of %id is more pronounced if local alignments are detected with highly sensitive spaced seeds

[9]. Third, even if Mauve often achieves higher coverage than our method YH, the reliability of some of its regions aligned in the fourth phase is questionable, and their %id argues in favour of discarding them from the output (see the *P.marinus* example).

With its publication, Mauve opened the way to a better handling of rearrangements; nonetheless our results suggest the similarity detection could be improved, and thereby the global reliability of the complete alignment. The comparison of the coverage of known orthologs between MGA, Mauve, and YH corroborates these findings. Interestingly, our program YH performs drastic improvements where both MGA and Mauve fail: on species with highly divergent strains like *B. aphidicola*, *P.marinus*.

Besides this gain in coverage and percent identities over MGA or sometimes Mauve, YH runs faster (a maximum running time of 102 s. and an average of 10 s.) and brings qualitative ameliorations. Its chain contains 150 anchors in average vs several thousands for MGA and Mauve, making it simpler to visualise and to grasp. Moreover, all local alignments it includes have an associated E-value that lies above a given threshold, ensuring they are statistically significant, which is not the case in MGA or Mauve alignments. Altogether, YH could be useful to automatically determine the backbone (a goal of Mosaic) without further post-processing based on an arbitrary threshold.

Acknowledgements: RU benefits from a PhD fellowship from the French Ministery of Research.This work is supported by the ANR project CoCoGen (BLAN07-1_185484).

References

[1] Bigot, S., Saleh, O., Lesterlin, C., Pages, C., Karoui, M.E., Dennis, C., Grigoriev, M., Allemand, J.F., Barre, F.X., Cornet, F.: KOPS: DNA motifs that control E. coli chromosome segregation by orienting the FtsK translocase. EMBO J. **24** (2005) 3770–3780

[2] Delsuc, F., Brinkmann, H., Philippe, H.: Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics **6** (2005) 361–375

[3] Hohl, M., Kurtz, S., Ohlebusch, E.: Efficient multiple genome alignment. Bioinformatics **18**(S1) (2002) S312–320

[4] Darling, A.C., Mau, B., Blattner, F.R., Nicole T. Perna: Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Res. **14**(7) (2004) 1394–1403

[5] Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I., Batzoglou, S.: Glocal alignment: finding rearrangements during alignment. Bioinformatics **19**(S1) (2003) i54–62

[6] Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.: Versatile and open software for comparing large genomes. Genome Biology **5**(2) (2004) R12

[7] Abouelhoda, M.I., Ohlebusch, E.: Chaining algorithms for multiple genome comparison. J. of Discrete Algorithms **3** (2005) 321–341

[8] Blanchette, M., Bourque, G., Sankoff, D.: Breakpoint phylogenies. In Miyano, S., Takagi, T., eds.: Genome Informatics. (1997) 25–34

[9] Noe, L., Kucherov, G.: YASS: enhancing the sensitivity of DNA similarity search. Nucl. Acids Res. **33**(S2) (2005) W540–543

[10] Roytberg, M.A., Ogurtsov, A.Y., Shabalina, S.A., Kondrashov, A.S.: A hierarchical approach to aligning collinear regions of genomes. Bioinformatics **18**(12) (2002) 1673–1680

[11] Sterk, P., Kersey, P.J., Apweiler, R.: Genome Reviews: Standardizing Content and Representation of Information about Complete Genomes. OMICS: A Journal of Integrative Biology **10**(2) (2006) 114–118

[12] Chiapello, H., Bourgait, I., Sourivong, F., Heuclin, G., Gendrault-Jacquemard, A., Petit, M.A., El Karoui, M.: Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. BMC Bioinformatics **6**(1) (2005) 171

[13] Schneider, A., Dessimoz, C., Gonnet, G.H.: OMA Browser Exploring orthologous relations across 352 complete genomes. Bioinformatics **23**(16) (2007) 2180–2182

Drug dosage control of the HIV infection dynamics

Marie-José MHAWEJ[1], Claude H. Moog[1]

Institut de Recherche en Communications et Cybernétique de Nantes, UMR CNRS 6597
1, rue de la Nöe BP 92101 44321 Nantes Cedex 3 France
Marie-Jose.Mhawej;claude.moog@irccyn.ec-nantes.fr

Abstract: *An increasing number of sophisticated control algorithms become available in the current literature to optimize the HIV therapy. Unfortunately, the pharmacokinetics and pharmacodynamics of antiretroviral drugs are ignored and these algorithms remain purely theoretic. This issue is investigated explicitly in this paper. An elementary pharmacodynamics model is combined with a non linear feedback control computed from standard engineering methods. It is shown that it results in the design of a realistic dosage regimen which drives the immunological system close to the healthy equilibrium state. Although the problem is dealt as a single input system, it is argued that the procedure can be extended to a multitherapy design or to any available control law.*

Keywords: HIV, nonlinear systems, input-output linearization, pharmacokinetics, pharmacodynamics

1 Introduction

Les modèles mathématiques de l'infection par le VIH qui existent de nos jours décrivent les dynamiques des lymphocytes T-CD4+ sains (principales cibles du virus), des lymphocytes T-CD4+ infectés, de la charge virale et parfois des cellules CD8+. Les premiers travaux en ce sens datent des années 90 [1,2,6,7]. Ils ont permis par exemple d'estimer les durées de vie *in vivo* du virus et des cellules infectées [7,6]. Des travaux plus récents mettent l'accent sur l'identification des paramètres des modèles mathématiques [10,11,8] ou encore sur l'application des théories de la commande pour l'optimisation des traitements antirétroviraux [3,4,5].

Les multi-thérapies qui existent actuellement se composent essentiellement des Inhibiteurs de la Protéase (PI) qui perturbent la maturation des nouveaux virions et des Inhibiteurs de la Transcriptase Inverse (RTI) qui empêchent la production de nouveaux virions en bloquant la transcription inverse de l'ARN viral en ADN. D'un point de vue d'automaticien, les traitements sont les entrées du système.

Nous introduisons ici des concepts de pharmacocinétique (PK) et de pharmacodynamique (PD) des antirétroviraux (ARV) qui permettent d'affiner la modélisation de l'entrée de commande, aspect jusque là négligé dans les travaux de commande. Nous illustrons ce travail en dérivant un régime thérapeutique « réaliste » basé sur la commande par linéarisation entrée-sortie du modèle mono-entrée. L'organisation de ce papier est comme suit : en Section 2, nous reprenons le modèle de base de la dynamique de l'infection par le VIH. Une loi de commande par linéarisation partielle est calculée en Section 3. La Section 4 introduit les principes de base de la pharmacocinétique et de la pharmacodynamique utilisés en Section 5. Cette dernière présente la contribution majeure de ce papier, qui consiste à calculer une posologie optimale en tenant compte de la pharmacologie des ARV. Enfin, la Section 6 conclut.

2 Modélisation

Nous présentons dans cette section le modèle de base décrivant la dynamique de l'infection par le VIH. Ce modèle à trois dimensions introduit dans [7], fait intervenir trois grandeurs caractéristiques : la population de cellules CD4+ saines (T) en $(CD4/mm^3)$, la population de cellules CD4+ infectées (T^*) en $(CD4/mm^3)$ et la charge virale (V) en $(copies d'ARN/ml)$.

$$\begin{cases} \dot{T} = s - \delta T - \beta TV, \\ \dot{T^*} = \beta TV - \mu T^*, \\ \dot{V} = kT^* - cV. \end{cases} \tag{1}$$

Ce modèle suppose que les CD4+ sains sont produits à un taux constant s et meurent à un taux δ. Ils sont infectés à la « vitesse » βTV proportionnelle à leur nombre et à la charge virale (V). Les CD4+ infectés meurent à un taux μ et les virus à un taux c. Les virions sont produits par les CD4+ infectés à un taux kT^*. Une analyse des données cliniques [9] indique que le traitement influe essentiellement sur le paramètre k. Cette étude n'est pas reprise dans ce papier faute de place. Cependant, nous en retenons le résultat principal : le modèle 3D est un modèle mono-entrée qui s'écrit :

$$\begin{cases} \dot{T} = s - \delta T - \beta TV, \\ \dot{T^*} = \beta TV - \mu T^*, \\ \dot{V} = (1 - u(t))kT^* - cV. \end{cases} \tag{2}$$

$u(t)$ $(0 \leq u(t) \leq 1)$ est l'entrée de commande unique affectant le paramètre k. $u(t)$ représente l'efficacité globale de la thérapie, *i.e* la superposition des effets des RTI et des PI.

3 Commande du modèle mono-entrée

3.1 Linéarisation entée-sortie

La linéarisation entrée-sortie est une technique standard dans la commande des systèmes non linéaires. Soit le système non linéaire :

$$\begin{aligned} \dot{x} &= f(x) + g(x).u \\ y &= h(x) \end{aligned} \tag{3}$$

Selon la théorie de commande non linéaire, l'approche consiste à trouver un changement de coordonnées (diffeomorphisme) $z = P(x)$ et un retour d'état $u(t) = a(x) + b(x)v$ qui transforme le système non linéaire entrée-sortie en un système linéaire équivalent. Après cette transformation, un régulateur est calculé pour le modèle linéaire obtenu.

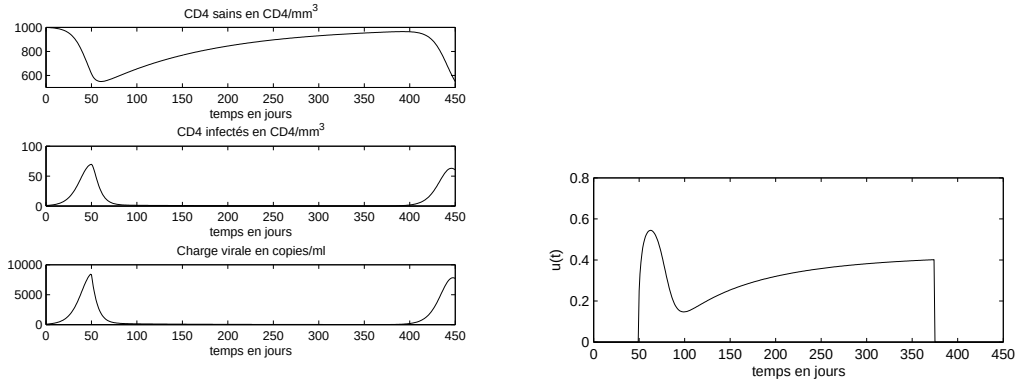
3.2 Application au modèle (2)

La commande $u(t)$ devrait ramener le système au point d'équilibre $(T_0 = s/\delta,\ T_0^* = 0,\ V_0 = 0)$. En choisissant $y = h(x) = (T - T_0) + T^*$, le système a un degré relatif [1] égal à 3 et est complètement linéarisable. Nous calculons alors le difféomorphisme $z = P(x)$ ainsi que l'expression

[1] le degré relatif d'un système est le nombre de fois qu'il faut dériver la sortie y pour obtenir une dépendance explicite de l'entrée u.

de la commande linéarisante $u(t)$. Ces expressions ne sont pas explicitées dans ce papier par manque de place, mais le lecteur peut se référer à [9] pour plus de détails. Cependant, la simulation de la linéarisation complète nécessite la manipulation d'expressions assez compliquées. Nous avons alors choisi la linéarisation partielle du système avec une dynamique de zéro stable, commande qui assure la convergence asymptotique vers le point d'équilibre désiré. Notons que le choix de la linéarisation partielle est aussi légitime que tout autre choix, celui-ci pouvant être aussi une commande linéaire fondée sur l'approximation linéaire autour d'un point d'équilibre [8].

Soit $y = h(x) = T - T_0$ la sortie linéarisante. Dans ce cas, le système a un degré relatif 2 et est partiellement linéarisable. Les expressions du changement de coordonnées partiel, de la commande linéarisante et de la dynamique de zéro dans le cas de la linéarisation partielle ne sont pas décrits ici mais se trouvent aussi dans [9]. La figure 1(a) est une simulation de la linéarisation partielle après un placement de pôles convenable. La commande $u(t)$ est représentée sur la figure 1(b). Le traitement est administré entre les jours 50 et 375 ($u(t) \neq 0\ pour\ 50 \leq t \leq 375$). Les paramètres et les conditions initiales de simulation sont : $s = 9\ mm^{-3}d^{-1}$, $\delta = 0.009\ d^{-1}$, $\beta = 4.1 \times 10^{-6}\ mld^{-1}$, $\mu = 0.3\ d{-}1$, $k = 75\ mm^3ml^{-1}d^{-1}$, $c = 0.6\ d^{-1}$, $c_1 = -0.0013$, $c_2 = -0.17$, $T(0) = 1000\ CD4/mm^3$, $T^*(0) = 1\ CD4/mm3$, $V(0) = 50\ copies/ml$.



(a) L'évolution de l'infection (b) L'entrée de commande

Fig. 1. Commande de la dynamique de l'infection par linéarisation entrée-sortie

Un avantage majeur de cette loi de commande est sa capacité à conduire le système vers le point d'équilibre désiré par le biais d'une efficacité de traitement variable au cours du temps. En d'autres termes, si l'efficacité du traitement est liée à la dose administrée, on peut dire qu'une dose complète n'est pas nécessaire tout le temps (voir Section 5).

4 Notions de pharmacologie

La pharmacocinétique (PK) d'un médicament est définie comme étant la relation qui existe entre la posologie de ce médicament et sa concentration dans le plasma alors que la pharmacodynamique (PD) représente la relation entre la concentration plasmique du médicament et son effet final.

1. **Modèle pharmacocinétique**
   Souvent, les médicaments sont administrés à doses constantes et à intervalle de temps constant. En négligeant le phénomène d'absorption, la quantité $X(t)$ de médicament dans le corps est régie par l'équation différentielle du 1^{er} ordre suivante :

$$\dot{X} = -KX\ pour\ n\tau < t < (n+1)\tau\ (n \in \mathbb{N}) \tag{4}$$

où K est la constante de temps d'élimination du médicament du 1^{er} ordre. Au début de chaque intervalle de dosage, la condition initiale s'écrit : $X_0(n\tau^+) = X(n\tau^-) + D_0(n\tau)$ où $D_0(n\tau)$ est la dose administrée à l'instant $n\tau$. Ce modèle à un compartiment est le modèle pharmacocinétique le plus simple présenté dans [12,13].

2. **Modèle pharmacodynamique**
La relation entre la réponse thérapeutique ou l'efficacité du médicament ($\eta(t)$) et la concentration plasmique $C(t)$ est empiriquement approximée par ([13,14]) : $\eta(t) = \eta_{max}\frac{C(t)^s}{C(t)^s+1/Q}$.
Souvent, $s = 1$, $\eta_{max} = 1$ et $1/Q = C_{50}$, concentration plasmique pour laquelle le médicament est efficace à 50%. Ainsi, en termes de quantité de médicament dans le corps, nous avons :

$$\eta(t) = \frac{X(t)}{X(t) + X_{50}} \tag{5}$$

où $X(t) = C(t) \times V_d$, $X_{50} = C_{50} \times V_d$, et V_d le volume de distribution du médicament.

5 Calcul de posologie

5.1 Le principe de la méthode

Cette méthode consiste en deux étapes :

1. Calcul de la commande nominale continue du modèle (2) selon les diverses techniques de commande non linéaire telles que la linéarisation entrée-sortie,
2. Calcul de la posologie échantillonnée équivalente à la posologie continue. Cette équivalence est fondée sur la modélisation pharmacologique et est détaillée ci-dessous.

Reprenons la commande de la section 3.2 Figure 1(b), et supposons que le patient est sous monothérapie. Alors, $u(t) = \eta_{désiré}(t)$. D'après l'équation (5), on peut calculer la quantité désirée de médicament dans le corps à chaque instant t : $X_{désiré}(t) = X_{50}\frac{\eta_{désiré}(t)}{1-\eta_{désiré}(t)}$. Sur une période interdose de longueur τ donnée, la quantité de médicament dans le corps qu'il faut est $\int_{n\tau}^{(n+1)\tau} X_{désiré}dt$. Nous calculons alors la quantité de médicament $X_0(n\tau)$ qu'il faut avoir au début de chaque intervalle de dosage de sorte à garantir la quantité totale de médicament dans le corps calculée précédemment sur cet intervalle. Selon le modèle pharmacocinétique (4), nous devons avoir :

$$\int_{n\tau}^{(n+1)\tau} X_0(n\tau)e^{-K(t-n\tau)}dt = \int_{n\tau}^{(n+1)\tau} X_{désiré}dt \tag{6}$$

$$X_0(n\tau) = \frac{K}{1-e^{-K\tau}}\int_{n\tau}^{(n+1)\tau} X_{désiré}dt \tag{7}$$

Les doses $D_0(n\tau)$ sont données par

$$D_0(n\tau) = X_0(n\tau) - X(n\tau^-) = X_0(n\tau) - X_0((n-1)\tau)e^{-K\tau} \approx X_0(n\tau).$$

La quantité $X(t)$ de médicament réellement présente dans le corps est la réponse du système du premier ordre (4) au train d'impulsions $X_0(n\tau)$. L'équation (5) permet de calculer l'efficacité réelle $\eta(t)$ du traitement. Nous discutons dans la section 6 la validité du principe pour les multithérapies.

5.2 Le cas de la monothérapie

Considérons le cas de la monothérapie à la zidovudine (ZDV). Les paramètres *in vivo* de ce médicament sont donnés dans [14] : $K = 0.35h^{-1}$ et $C_{50} = 0.8\ mg/L$. D'après [15], le volume de distribution par kg de la zidovudine est $v_d = 1.6L/kg$. Pour une personne de masse corporelle moyenne $M = 70kg$, on a $X_{50} = C_{50} \times v_d \times M = 89,6mg$.
La Figure 2 montre la posologie de ZDV calculée selon la méthode décrite à la Section 5.1 pour $\tau = 0,5\ jour$. On remarque sur cette figure que les dosages calculés sont compris entre 61.9 mg et 434.5 mg pour une moyenne de 190.3 mg. Ces valeurs sont en cohérence avec la dose administée en trithérapie standard, soit 300 mg b.i.d. D'autre part, la Figure 3 illustre l'évolution de l'infection (charge virale, figure 3(a) et CD4 sains, figure 3(b)) pour une monothérapie à la ZDV une fois par jour ($\tau = 1jour$), deux fois par jour ($\tau = 0,5jour$) et trois fois par jour ($\tau = 0.3jour$). Nous rappelons qu'à chaque fois, la posologie ($X_0(n\tau)$) est calculée selon le principe de la Section 5.1. Sur cette figure, nous pouvons voir que plus τ diminue, plus la réponse à la thérapie s'approche de la réponse optimale et du point d'équilibre désiré. Ceci justifie l'échec de la prise de médicament une fois par jour face à des administrations quotidiennes multiples.

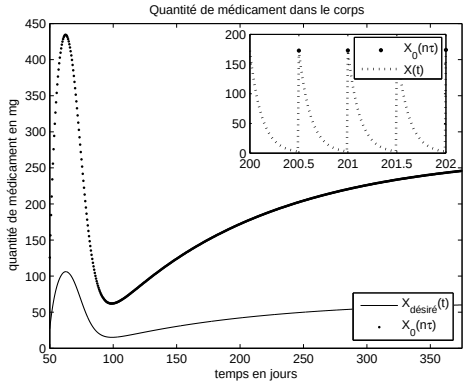
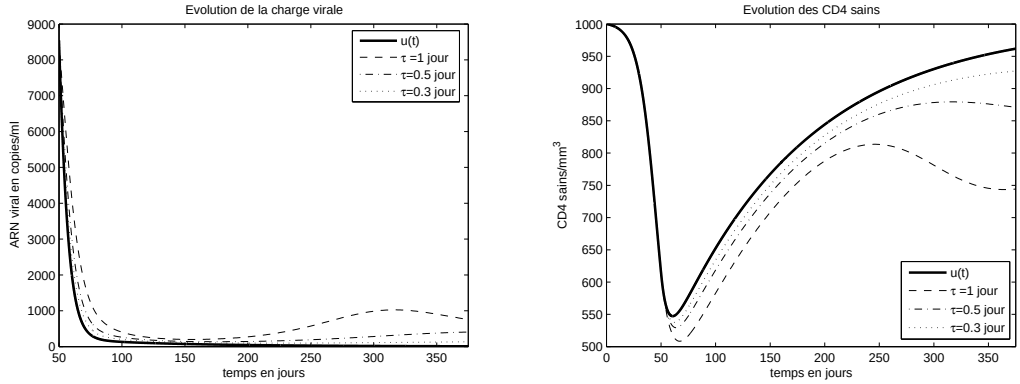


Fig. 2. Posologie calculée pour $\tau = 0,5jour$



(a) Evolution de la charge virale (b) Evolution des CD4 sains

Fig. 3. Evolution de l'infection pour différentes valeurs de τ

6 Conclusion et perspectives

Dans ce papier, une loi de commande de linéarisation par bouclage est calculée pour le modèle 3D de la dynamique de l'infection par le VIH. Cette loi ramène le système vers l'état d'équilibre désiré, à savoir, un taux de CD4 sains voisin de $1000\ CD4/mm^3$ et une charge virale indétectable. Par ailleurs, l'incorporation des modèles PK et PD des ARV ainsi que la « discrétisation » de la loi de commande selon le principe décrit à la Section 5.1 ont permis de déduire pour une première fois un régime thérapeutique pragmatique en termes de posologie.

La suite de ce travail consiste à traiter le cas de la multithérapie. Pour cela, si p médicaments sont administrés nous prenons $u(t) = \eta_{désiré}(t) = 1 - \prod_{i=1}^{p}(1 - \eta_{i\ désiré})$. Il suffit alors de poser $(p-1)$ autres équations indépendantes de la forme $f(\eta_{1\ désiré}, \eta_{2\ désiré}, \ldots, \eta_{i\ désiré}, \ldots) = 0$ pour calculer les $\eta_{i\ désiré}$. Nous revenons ensuite au cas de p monothérapies.

Remerciement

Nous remercions les patients qui ont consenti à participer à cette étude dont le CHU de Nantes était promoteur.

Références

[1] Ho, David D. *et al.*, Rapid turnover of plasma virion and CD4 lymphocytes in HIV-1 infection, Nature, (373) (1995), 123-126.

[2] Wei, X. *et al.*, Viral dynamics in Human Immunodeficiency virus type 1 infection, Nature, (373) Jan. 1995, 117-122.

[3] D. Kirschner, S. Lenhart, S. Serbin, Optimal control of the chemotherapy of HIV, J. Math. Biol 35 (1997) 775-792.

[4] H. Shim, S.-J. Han, H. Shim, S. Nam, J. Seo, Optimal scheduling of drug treatment for HIV infection : Continuous dose control and receding horizon control, International Journal of Control, Automation, and Systems 1 (3).

[5] H. Chang, A. Astolfi, Control of HIV infection dynamics, IEEE Control systems magazine (2008) 28-39.

[6] Perelson, A.S. *et al.*, Decay characteristics of HIV-1 infected compartment during combination therapy, Nature, (387) (1997), 188-191.

[7] A. Perelson, P. Nelson, Mathematical analysis of HIV-1 dynamics in vivo, SIAM Review 41 (1) (1999) 3-44.

[8] D. Ouattara, Modélisation de l'infection par le VIH, identification et aide au diagnostic, Ph.D. thesis, Ecole Centrale de Nantes & Université de Nantes, Nantes, France (Sep. 2006).

[9] M.J. Mhawej, C.H. Moog, F. Biafore, Control of the HIV infection and drug dosage, Submitted to Biomedical Signal Processing and Control, 2008.

[10] R. Filter, X. Xia, A penalty function to HIV/AIDS model parameter estimation, in : 13th IFAC Symposium on System Identification, Rotterdam, 2003.

[11] D. Ouattara, Mathematical analysis of the HIV-1 infection : Parameter estimation, therapies effectiveness, and therapeutical failures, in : 27th Annual International Conference of the IEEE Engineering in Medecine and Biology Society, Shanghai, China, 2005.

[12] M. Gibaldi, D. Perrier, Drugs and the pharmaceutical sciences, Pharmacokinetics, Vol. 1, Marcel Dekker, 270 Madison Avenue, New York, New York, 10016, 1975.

[13] G. Wagner, Fundamentals of clinical pharmacokinetics, Drug Intelligence Publications, inc, 1975.

[14] M. Legrand, E. Comets, G. Aymard, R. Tubiana, C. Katlama, B. Diquet, An in vivo pharmacokinetic/pharmacodynamic model for for antiretroviral combination, HIV Clinical trials 4 (3) (2003) 170-183.

[15] HIV pharmacology, available at http ://www.hivpharmacology.com/.

Détection de nouveaux domaines protéiques par co-occurrence : application à *Plasmodium falciparum*

Nicolas Terrapon[1,2], Olivier Gascuel[1], Laurent Bréhélin[1]

[1] LIRMM, Univ. Montpellier 2, CNRS, 161 rue Ada 34392 Montpellier Cedex 5 France
[2] CEA Grenoble iRTSV/LPCV, 17 rue des Martyrs, 38054 Grenoble cedex 9 France

Abstract: *Hidden Markov Models (HMMs) have proved to be powerful for protein domain identification. However, numerous domains may be missed in highly divergent proteins. This is the case for the proteins of Plasmodium falciparum, the main causal agent of human malaria. Here, we propose a method that uses domain co-occurrence to increase the sensitivity of the approach while controlling its false discovery rate. Applied to P. falciparum, our method identify (with an error rate below 20%) 590 new domains (versus 3683 in Pfam Database), which involve 283 new GO annotations.*

Keywords: Hidden Markov Models, Protein Domains, Gene Ontology, Malaria.

1 Introduction

Les *modèles de Markov cachés* (HMM [1]) se sont révélés être un outil puissant pour l'identification de domaines protéiques grâce à leur capacité à capturer l'information spécifique à chaque position. Chaque HMM représente un domaine donné. Étant donné une nouvelle séquence protéique, l'approche probabiliste permet de calculer un score qui reflète la probabilité que le HMM ait généré la séquence. Ce score peut aussi être utilisé pour calculer une E-valeur, espérance du nombre de séquences ayant un aussi bon score dans une base de séquences aléatoires. La base de données en ligne Pfam (version 23.0) [2] propose une large collection de HMM modélisant des familles de domaines couvrant plus de 73% des protéines d'Uniprot [3]. Un certain nombre de domaines Pfam sont annotés dans la *Gene Ontology* ou GO [4]. L'annotation d'un domaine correspond aux informations communes à toutes les protéines ayant ce domaine [5], ce qui permet, lorsqu'un nouveau domaine est identifié dans une protéine, de transférer les annotations GO du domaine à la protéine. Pfam fournit avec ses modèles des seuils permettant d'affirmer la présence du domaine si le score de la séquence est supérieur au seuil. Cependant, chez certaines protéines fortement divergentes, cette approche n'est pas assez sensible pour permettre l'identification des domaines composants la protéine. Appliquée à *P. falciparum* par exemple (l'agent responsable de la forme létale de la malaria humaine), cette stratégie se révèle incapable de détecter le moindre domaine dans 47% de ses protéines, tandis que de nombreux domaines semblent absents du répertoire de *P. falciparum* (seulement 1421 domaines distincts ont pu être identifiés). À titre de comparaison 2369 domaines sont répertoriés chez la levure, et concernent 76% des protéines. Une des explications à ces difficultés réside dans le fort biais compositionnel des protéines de *P. falciparum*, induit par la composition à 80% de A+T de son génome. Relâcher les seuils requis pour la détection des domaines permettrait de plus nombreuses annotations, mais au prix d'un nombre d'erreurs important. Une solution est alors d'utiliser des informations supplémentaires pour filtrer parmi ces domaines potentiels ceux qui ont le plus de chance d'être réellement présents. Dans cet article nous proposons d'utiliser la co-occurrence de domaines pour cela.

Les différentes études publiées concernant la combinatoire des compositions en domaines des protéines révèlent un certains nombre de propriétés. Les protéines composées des mêmes domaines ont généralement une fonction similaire [6]. La conservation de groupes de domaines au cours de l'évolution a été mise en évidence par plusieurs études montrant que le nombre de combinaisons de domaines identifiés dans la nature est infime en comparaison du nombre de combinaisons possibles : les domaines protéiques n'apparaissent qu'avec un nombre nombre limité d'autres domaines favoris au sein des protéines [7].

Nous présentons dans un premier temps notre méthode de recherche par co-occurrence ainsi qu'une procédure permettant de contrôler le taux d'erreur de la méthode. Nous validons ensuite notre approche grâce à des simulations sur la levure, puis nous présentons les résultats obtenus lorsqu'elle est appliquée à un organisme fortement biaisé comme *P. falciparum*.

2 Méthode

Nous proposons d'utiliser les propriétés de co-occurrence des domaines pour *certifier* la présence d'un domaine potentiellement présent dans une protéine à partir de la présence avérée d'un autre domaine. Notre approche consiste dans un premier temps à identifier parmi toutes les protéines de Uniprot, les paires de domaines montrant une co-occurrence forte (vérifiée par un test statistique) dans de nombreuses protéines. Ces paires de domaines conditionnellement dépendants (*PDCD*) forment alors une liste de référence qui est utilisée de la manière suivante. Considérons une protéine de notre organisme cible (par exemple *P. falciparum*) pour laquelle un ou plusieurs domaines sont déjà connus. En relâchant les seuils de score, les HMM de Pfam détectent un ou plusieurs nouveaux domaines potentiels. Si l'un de ces domaines forme, avec au moins un des domaines connus de la protéine, une paire faisant partie de la liste des *PDCD* de référence alors il est considéré comme certifié. Pour appliquer cette méthode de certification par co-occurrence, on a donc besoin de connaître, pour chaque protéine i de l'organisme étudié, l'ensemble de ses domaines *avérés* (A_i) et *potentiels* (P_i). Il faut aussi établir à l'aide de l'ensemble des protéines de composition connue, la liste de paires de domaines co-occurrents de référence, notée *PDCD* qui permet de certifier un domaine potentiel $x \in P_i$, grâce à un domaine avéré $y \in A_i$, si $(x, y) \in$ *PDCD*.

L'ensemble des domaines potentiels (P_i) se construit à partir des résultats de la recherche des HMM de Pfam sur la séquence protéique i grâce au logiciel `hmmer` [8]. Elle est paramétrée pour fournir l'ensemble des domaines dont l'E-valeur est inférieure à une valeur beaucoup moins stringente que la valeur seuil proposée par Pfam. Les résultats sont ensuite traités pour obtenir un ensemble de domaines non-recouvrants. Cette opération est effectuée grâce à un algorithme de pavage qui conserve en priorité les domaines possédant la meilleure E-valeur. À l'issue de cette phase, on conserve pour chaque protéine i l'ensemble des domaines potentiels non redondants P_i.

La base de connaissance des domaines avérés (A_i) peut être construite de différentes manières. La plus sûre est d'extraire directement des bases de données dédiées aux organismes, les domaines Pfam dont la présence a été certifiée par des experts, par exemple la base PlasmoDB [9] (version 5.5) pour *P. falciparum*. Elle peut aussi être obtenue en effectuant une recherche à l'aide des HMM de Pfam sur l'organisme cible en respectant les seuils proposés par Pfam. Cependant, d'autres bases de connaissance complémentaires peuvent être envisagées. On peut par exemple s'appuyer sur l'ensemble des domaines Interpro [5] répertoriés dans notre organisme cible (issus de PlasmoDB pour *P. falciparum*). L'utilisation de l'intégralité des bases de données d'Interpro permet alors de disposer d'informations issues de 9 bases de domaines protéiques supplémentaires (SMART, PROSITE, Gene3D, Superfamily, PANTHER, Tigrfams, PRINTS, PIRSF, ProDom). En étendant de cette manière notre base de connaissances et en apprenant une liste de *PDCD* spécifique où chaque paire est composée d'un do-

maine Pfam et d'un domaine Interpro (non-Pfam), nous espérons pouvoir certifier plus de domaines, même dans des protéines où aucun domaine Pfam n'est connu. Néanmoins, comme pour la base Pfam, cette base limite la certification par co-occurrence à des protéines où au moins un domaine est déjà connu. Une autre base de connaissance complémentaire est de considérer les domaines potentiels (P_i) eux-mêmes comme base de connaissance. Dans cette solution, on essaye de certifier un domaine potentiel par un autre domaine potentiel (au risque d'un taux d'erreur plus important) afin de détecter des domaines Pfam dans des protéines où aucun domaine n'est connu.

La liste des paires de domaines conditionnellement dépendants est calculée à partir de l'ensemble des paires qui ont déjà été observées dans les protéines d'Uniprot chez d'autres organismes. Ces paires étant utilisées pour certifier la présence potentielle d'un domaine grâce à un autre domaine, elles doivent révéler une dépendance conditionnelle entre ces domaines, *i.e.* la présence de l'un doit être un indice fort de la présence de l'autre. Toutes les paires observées dans Uniprot ne satisfont pas ce critère. Par exemple, si deux domaines fréquents apparaissent avec de nombreux domaines différents (très versatiles), ils ne forment pas une paire conditionnellement dépendante. Tester la dépendance conditionnelle des paires de domaines revient à mesurer l'association de deux variables. On doit effectuer un test de comparaison entre deux proportions correspondant à l'observation simultanée de deux caractères différents sur les mêmes individus. Les individus sont les N protéines multidomaines d'Uniprot dont la composition en domaines est connue. Les deux caractères observés dans ces protéines sont la présence (ou l'absence) des domaines formant chaque paire. Une solution à ce problème peut être apportée par un test de corrélation de type χ^2. Nous avons choisi d'appliquer un test exact de Fisher, plus adapté pour de petits échantillons comme c'est le cas ici. Pour chaque paire de domaines une P-valeur peut donc être calculée. Si cette P-valeur est inférieure à un certain seuil (typiquement 1%) l'hypothèse nulle est rejetée, les domaines sont considérés comme conditionnellement dépendants, et la paire est ajoutée à la liste des *PDCD*.

Contrôle du taux de faux positifs : À partir des domaines potentiels, des domaines avérés et de la liste des *PDCD*, on est capable de certifier un certain nombre de domaines inédits. Une question est alors de pouvoir estimer le nombre de domaines certifiés par erreur par notre approche. Pour cela nous proposons d'estimer l'espérance du nombre de nouveaux domaines que notre approche certifierait sous l'hypothèse H_0 où tous les domaines potentiels étaient prédits de manière aléatoire. Cela peut être réalisé par simulation, à l'aide d'une procédure de permutation aléatoire des différents domaines potentiels des protéines. Permuter les différents domaines crée une situation dans laquelle les domaines potentiels sont indépendants des domaines avérés, tout en préservant la distribution de ces domaines, ainsi que la distribution du nombre de domaines potentiels et avérés par protéine. La procédure de permutation est la suivante. Dans un premier temps, l'ensemble des domaines avérés associés aux protéines est fixé. Puis on collecte l'ensemble des domaines potentiels de toutes les protéines, et on les redistribue aléatoirement à travers les différentes protéines en créant de nouveaux ensembles de domaines potentiels P_i^* de même taille que les ensembles P_i originaux. On applique ensuite notre méthode sur ces domaines potentiels, et on comptabilise le nombre de domaines potentiels qu'elle certifie. On réitère cette procédure un grand nombre de fois (typiquement 1000), et on moyenne les résultats. Ce nombre moyen de domaines certifiés sous l'hypothèse H_0 est comparé au nombre de certifications réalisées sur les données originales. Le taux de faux positifs (estimation du *False Discovery Rate*, ou *FDR*) de la méthode est estimé par le ratio :

$$\widehat{FDR} = \frac{\text{espérance du nombre de certification sous } H_0}{\text{nombre de domaines certifiés sur les données originales}}.$$

En jouant sur le seuil d'E-valeur utilisé pour définir les domaines potentiels, on peut donc, grâce à cette procédure, contrôler le *FDR* associé à nos prédictions.

3 Résultats

La première expérience réalisée consistait à nous assurer de la capacité de la méthode à trouver les domaines qui échappent aux seuils de Pfam à cause d'une dérive trop importante des séquences protéiques. Le principe est le suivant. Les HMM de Pfam sont utilisés avec leurs seuils de score pour déterminer l'ensemble des domaines de référence chez *S. cerevisiae*, organisme choisi pour la qualité de ses annotations. On fait ensuite subir aux séquences de la levure une évolution rapide vers la composition de *P. falciparum* à l'aide du programme `seqgen` [10]. Nous avons ainsi créé 4 jeux de séquences protéiques artificiels de divergence croissante (grâce à des taux t de substitution de 0.1, 0.25, 0.5 et 0.75, une matrice de substitution, *WAG*, et une distribution d'acides aminés cible : la distribution moyenne chez *P. falciparum*), sur lesquelles on applique la procédure suivante. Dans un premier temps, chaque HMM est utilisé avec son seuil de Pfam pour détecter les domaines présents. On s'attend à ce qu'un certain nombre de domaines de référence ne soient plus détectés à cause de la dérive des séquences. Dans un second temps, nous relâchons les seuils (à une E-valeur de 10) et appliquons la méthode de certification par co-occurrence en utilisant les domaines encore détectés par les seuils de Pfam comme base de connaissance. On espère ainsi retrouver une partie des domaines précédemment perdus. Les résultats sont présentés dans le tableau 1. Par exemple, pour $t = 0.5$, des 907 domaines perdus, 645 sont potentiellement retrouvable (*i.e.* sont présents dans une protéine pour laquelle au moins un autre domaine est encore détecté), et 491 sont effectivement retrouvés. De plus, 60 inédits (absents des domaines de référence) sont également détectés. Pour les taux de substitution élevés, on remarque que la proportion d'inédits parmi les domaines certifiés (*i.e.* $\frac{\text{Domaines inédits}}{\text{Domaines retrouvés} + \text{Domaines inédits}}$) est proche du taux d'erreur estimé par notre procédure, ce qui tend à valider cette procédure. Pour les taux bas, par contre, on remarque que la proportion d'inédits est sensiblement plus haut que le taux d'erreur estimé. Une question est alors de savoir si parmi ces inédits une certaine partie ne serait pas de "vrais" domaines non encore référencés chez la levure. Pour vérifier cette hypothèse, nous avons calculé parmi les domaines retrouvés qui possèdent une annotation GO, la proportion possédant une annotation non référencée chez la protéine (dernière colonne du tableau 1). On constate que la proportion de domaines ayant une annotation GO inédite est beaucoup plus basse que la proportion de domaines inédits, et plus proche de notre *FDR*. Les autres domaines (apportant des annotations déjà connues) qui constituent l'essentiel des domaines inédits, sont concordants avec les annotations connues de la protéine. Cela semble indiquer que les "vrais" inedits (apportant des annotations GO inédites) sont en effet rare, comme on peut s'y attendre chez la levure, et donc qu'une partie des domaines inédits ne sont pas des faux positifs mais des domaines réellement présents que nous certifions grâce à notre approche.

Taux substitution	Domaines de référence	Domaines perdus	Potentiellement retrouvable	Domaines retrouvés	Domaines inédits	*FDR* Estimé	Proportion nvx GO
0.1	2407	149	145	134	274	11.5%	15%
0.25	2407	346	301	265	171	9.2%	7.8%
0.5	2407	907	645	491	60	5.4%	3.1%
0.75	2407	1436	747	501	12	4%	0.3%

Tableau 1. Résultats sur la levure après évolution. "Taux substitution" indique le taux de divergence des séquences, "Domaines de référence" les domaines des protéines multidomaines de la levure originale, "Domaines perdus" correspond aux domaines non retrouvés par les seuils de Pfam sur les séquences divergentes, "Domaines retrouvés" les domaines perdus que l'on retrouve par notre méthode de certification, "Domaines inédits" le nombre de domaines inédits à l'ensemble de référence trouvé en plus par notre méthode, et "Proportion nvx GO" la proportion de domaines ayant une annotation GO inédite vis à vis de la protéine.

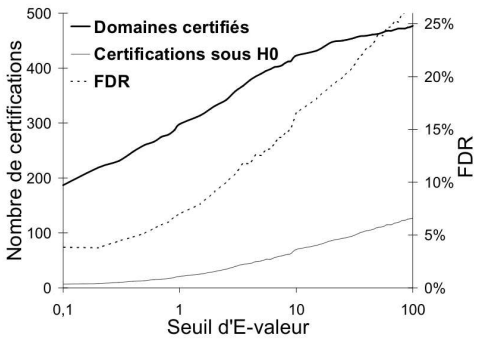


Fig. 1. Évolution du nombre de certifications, de l'estimation du nombre d'erreurs et du *FDR* en fonction de l'E-valeur (en abscisse). Le nombre de domaines certifiés (ligne épaisse) et le nombre d'erreurs estimées (ligne fine) évoluent en ordonnées sur l'axe de gauche, et le *FDR* (en pointillés) sur l'axe de droite.

Nous avons ensuite appliqué notre méthode à *P. falciparum* en utilisant les trois sources d'information détaillées en introduction : les domaines Pfam connus, les domaines Interpro non-Pfam connus, et les domaines potentiels eux-mêmes. La figure 1 présente les résultats obtenus pour différents seuils d'E-valeurs (en abscisse), en utilisant les domaines Pfam référencés dans PlasmoDB [9] et en utilisant une liste de *PDCD* selectionnée avec une P-valeur seuil de 1%. On constate comme attendu que le nombre de domaines certifiés ainsi que le *FDR* augmentent avec le seuil d'E-valeur utilisé pour la sélection des domaines potentiels. On peut donc, suivant que l'on désire un plus grand nombre de domaines certifiés ou un *FDR* faible, jouer sur le seuil d'E-valeur pour générer un ensemble de prédictions en accord avec l'objectif privilégié.

Le tableau 2 présente l'ensemble des résultats obtenus en utilisant les trois bases de connaissances pour différents seuils de *FDR* : les prédictions ayant un *FDR* inférieur à 10% et celles ayant un *FDR* inférieur à 20%. Par exemple, pour un *FDR* inférieur à 20%, 590 nouveaux domaines sont certifiés, parmi lesquels 516 correspondent à l'identification d'une nouvelle famille de domaines Interpro dans la protéine. Ils représentent un apport de 16% de domaines par rapport à l'ensemble des 3683 domaines Pfam connus chez *P. falciparum*. Les domaines Pfam connus permettent de certifier 406 nouveaux domaines, les domaines Interpro 329, et les domaines potentiels eux-mêmes 167 (avec du recouvrement, certains domaines étant certifiés par 2 ou 3 de ces bases). De plus, ces domaines certifiés avec un *FDR* inférieur à 20% ont permis la découverte de 191 types de domaines qui n'avaient jamais été observés dans une protéine de *P. falciparum* auparavant. Ces domaines vont s'ajouter au 1421 types de domaines connus chez *P. falciparum* (*cf.* section 1), soit une augmentation du nombre de domaines d'environ 13%. Enfin, parmi les nouveaux domaines certifiés chez *P. falciparum*, un certain nombre possèdent des annotations GO inédites qui peuvent être transférées aux protéines. Par exemple pour un *FDR* inférieur à 20%, les domaines certifiés apportent un total de 283 nouvelles annotations GO chez *P. falciparum* (soit 3,3% d'annotations supplémentaires si l'on se rapporte aux 8312 annotations GO de *P. falciparum*), 189 provenant d'un nouveau domaine ayant été certifié par co-occurrence avec des domaines Pfam connus, 109 avec des domaines Interpro connus et 68 avec les domaines potentiels. Nous avons notamment identifié dans la protéine MAL7P1.12, les domaines drsm et ResIII annotés par les termes GO *binding to double stranded RNA*, *DNA binding*, *ATP binding* et *hydrolase activity*, ce qui laisse supposer un rôle dans des processus cellulaires tels que la régulation/signalisation par ARN double-brin, ou un mécanisme de défense contre des pathogènes, ou encore un contrôle des niveaux d'ARN de la cellule du parasite au cours de son cycle de vie. Il serait important de préciser ce rôle compte tenu des débats actuels concernant la régulation des gènes de *P. falciparum* par l'ARN (phénomènes de *RNA decay*) plus que par des facteurs de transcription.

FDR	<10%				<20%			
Base connaissance	Pfam	Interp.	Pot.	Toutes	Pfam	Interp.	Pot.	Toutes
Domaines certifiés	298	191	109	400	406	329	167	590
Nvlles Familles Interpro	246	160	101	337	349	282	155	516
Domaines inédits chez *Pf*	99	67	47	130	139	103	64	191
Nvlles annotations GO	106	50	35	145	189	109	68	283

Tableau 2. Tableau récapitulatif des résultats sur *P. falciparum* pour différents tranches de *FDR*. "Base connaissance" correspond aux bases de connaissance des domaines avérés utilisées pour la certification : "Pfam", "Interp." pour Interpro, "Pot." pour les domaines potentiels et "Toutes" pour les résultats cumulés des trois bases. "Nvlles annotations GO" indique le nombre de nouveaux termes GO transférés aux protéines.

4 Conclusion

Nous avons présenté une méthode améliorant la sensibilité de la détection de domaines protéiques par des modèles probabilistes, en s'appuyant sur les propriétés de co-occurrence des domaines. Cette méthode qui a été initialement développée pour l'étude d'organismes dont l'annotation est pauvre (dûe à un protéome à fort biais compositionnel), peut aussi s'appliquer à des organismes déjà bien annotés. Nos résultats montrent qu'elle permet de certifier un nombre important de domaines, tout en contrôlant le taux d'erreur en fonction de l'objectif privilégié (nombreux nouveaux domaines ou *FDR* stringent). Appliquée à *P. falciparum*, elle permet par exemple de certifier 590 nouveaux domaines avec un *FDR* inférieur à 20% et d'apporter 283 nouvelles annotations GO à ses protéines.

Remerciements

Ce travail est soutenu par le projet ANR PlasmoExplore (ANR-06-MDCA-014). Nous remercions tout particulièrement Éric Maréchal, ainsi que l'ensemble des membres du projet PlasmoExplore.

References

[1] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press, New York, 1998.
[2] R.D. Finn, J. Tate, J. Mistry, P.C. Coggill, S.J. Sammut, H.R. Hotz, G. Ceric, K. Forslund, S.R. Eddy, E.L.L. Sonnhammer and A. Bateman, The Pfam Protein Families Database. *NAR*, 36:D281-D288, 2008.
[3] The UniProt Consortium, The Universal Protein Resource (UniProt). *NAR*, 36:D190-D195, 2008.
[4] The Gene Ontology Consortium, The Gene Ontology (GO) project in 2006. *NAR*, 34(Database issue):D322-D326, 2006.
[5] N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, D. Barrell, A. Bateman, *et al.*, The InterPro Database, 2003 brings increased coverage and new features. *NAR*, 31(1):315-318, 2003.
[6] M. Gerstein and H. Hegyi, Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Research*, 11:1632-1640, 2001.
[7] I. Cohen-Gihon, R. Nussinov and R. Sharan, Comprehensive analysis of co-occuring domain sets in yeast proteins, *BMC Genomics*, 8:161, 2007.
[8] S.R. Eddy, Profile Hidden Markov Models. *Bioinformatics*, 14:755-763, 1998.
[9] A. Bahl, B. Brunk, J. Crabtree, D. Gupta, J.C. Kissinger, D.S. Pearson, D.S. Roos DS, J. Schug, C.J. Jr Stoeckert *et al.*, PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *NAR*, 31(1):212-215, 2003.
[10] A. Rambaut and N.C. Grassly, Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 13:235-238, 1997.

Système de Classes Chevauchantes pour la Recherche de Protéines Multifonctionnelles.

Emmanuelle Becker[1], Alain Guénoche[2], Christine Brun[1]

[1] TAGC U928, INSERM / Université de la Méditerranée, Marseille, France
becker,brun@tagc.univ-mrs.fr
[2] IML UMR 6206, CNRS / Université de la Méditerranée, Marseille, France
guenoche@iml.univ-mrs.fr

Abstract: *This work aims at developing a method to detect multifunctional proteins, i.e. proteins performing several apparently unrelated functions. We consider a network of binary direct interactions between proteins that we decompose in an overlapping class system using a criteria based on graph topology and extending Newman's modularity. As a result, some proteins are found in several final classes meaning that they are interacting with several groups of proteins apparently functionally unrelated. These proteins are thus good candidates for multifunctionality. In this paper, we first introduce the concept of multifunctionality, then explain the method, and finally present the preliminary results obtained by applying the method to a large human protein interaction network.*

Keywords: Overlapping class system, protein interaction network, moonlighting protein.

1 Introduction

Certaines protéines assurent des fonctions très différentes dans la cellule. Par exemple, la protéine alpha-cristalline humaine est à la fois un composant structural du cristallin et elle est impliquée dans la réponse au choc thermique lorsqu'elle est exprimée dans d'autres tissus. Ces protéines qualifiées de multifonctionnelles ou "moonlighting proteins" (to moonlight = cumuler deux emplois) peuvent, de par leur singularité fonctionnelle, jouer des rôles régulateurs importants, permettre de comprendre la complexité de certains phénotypes ou les effets secondaires de certaines drogues [1]. Cependant leur découverte ne s'est faite, jusqu'à présent, que fortuitement : les approches expérimentales déterminant la fonction des protéines étant dirigées par des hypothèses, leur identification en tant que protéines multifonctionnelles n'a pu se faire que par la convergence non anticipée de résultats expérimentaux. Il existe donc un besoin méthodologique pour l'identification de telles protéines qui soit sans *a priori* et à grande échelle.

Les réseaux d'interactions protéine-protéine (=IPP) correspondent à l'ensemble des interactions physiques détectées entre les protéines d'un organisme [2]. Leur analyse pourrait contenir des informations fonctionnelles pertinentes pour l'identification de protéines multifonctionnelles (=PMF). En effet, les protéines interagissant spécifiquement avec d'autres partenaires protéiques pour assurer leur fonction, il est attendu que des protéines ayant plusieurs fonctions interagissent avec des groupes de partenaires différents au sein du réseau, selon la fonction considérée. Comme les réseaux d'IPP sont représentés par des graphes simples dans lesquels les sommets correspondent aux protéines et les arêtes aux interactions directes, des méthodes de partitionnement de graphe d'IPP ont été proposées

ces dernières années pour la mise en évidence de groupes de sommets fortement connectés (pour une revue, voir [3]). Ces méthodes ont contribué à l'identification de modules fonctionnels regroupant des protéines impliquées dans la même voie ou le même processus cellulaire. Cependant, elles aboutissent à des partitions strictes du graphe n'autorisant pas l'affectation d'un sommet à plusieurs classes, alors que cette possibilité est nécessaire pour une PMF. En effet, étant impliquée dans des fonctions différentes, elle doit pouvoir participer à des modules fonctionnels différents.

Nous voulons donc rechercher les PMF dans les intersections de classes chevauchantes construites à partir des graphes d'IPP. Dans cet article, nous présentons dans un premier temps une méthode permettant de recouvrir un graphe par un système pertinent de classes chevauchantes. Cette méthode est ensuite appliquée à un réseau d'IPP humaines (24 000 interactions). Au sein des intersections de classes chevauchantes, nous recherchons alors un ensemble de PMFs identifiées à partir de l'analyse de la littérature et évaluons les performances de la méthode.

2 Méthode de Construction du Système de Classes Chevauchantes

Les systèmes de classes chevauchantes sont apparus dans les années 80 par le biais d'études théoriques sur des familles de distances (pour une revue, voir [4]). Hormis dans le cas des pyramides, très liées à l'existence d'un ordre total sur les sommets, l'application de ces modèles à des données réelles n'a pas connu le même succès que les méthodes hiérarchiques ou celles de partitionnement.

2.1 Choix du Critère : Modularité $M(P)$, Modularité étendue $Q(\alpha)$

L'objectif de déctecter des groupes de protéines densément connectés et patageant des fonctions communes est traduit mathématiquement par celui de fabriquer des classes ayant un grand nombre d'arêtes internes relativement aux cardinaux des classes. Parmi les critères récents introduits dans ce but, la notion de modularité, définie par Newman dans le cadre de partitions strictes [5], permet de quantifier le surcroît d'arêtes internes par rapport à ce que l'on obtiendrait avec une partition aléatoire du graphe ayant les mêmes cardinaux des classes.

Soit $G = (V, E)$ un graphe simple connexe à n sommets et m arêtes ($|V| = n, |E| = m$) et P une partition de V en p classes : $P = \{V_1, V_2, ..V_p\}$. Soit e_{ij} le pourcentage d'arêtes ayant une extrémité dans la classe V_i et l'autre dans la classe V_j : $e_{ij} = |E \cap (V_i \times V_j)|/m$. La probabilité qu'une arête tirée au hasard ait une extrémité dans la classe V_i est alors : $a_i = e_{ii} + 1/2 \sum_{j \neq i} e_{ij}$ et la modularité de la partition P est définie par :

$$M(P) = \sum_{i=1..p} (e_{ii} - a_i^2).$$

Très récemment, plusieurs auteurs ont établi des critères équivalents qui permettent d'étendre la modularité aux systèmes de classes, c'est à dire aux recouvrements par un ensemble de classes chevauchantes [7,8]. Soit d_x le degré du sommet x dans G et A sa matrice d'incidence ($A_{xy} = 1$ ssi $(x, y) \in E$). On note B la matrice de terme général $B_{xy} = 2mA_{xy} - d_x d_y$. Les valeurs de B associées aux arêtes de G ($A_{xy} = 1$) sont positives ou nulles, sauf si $d_x d_y > 2m$, et inversement toutes les valeurs de B correspondant aux paires de sommets non connectés sont négatives. On admettra dans la suite que toutes les arêtes sont à valeur positive dans B.

Un système de classes est défini par une relation binaire $\alpha : V \times V \rightarrow \{0, 1\}$, telle que $\alpha_{xy} = 1$ si les sommets x et y sont réunis dans au moins une classe et 0 sinon. Angelelli & Reboul [7] ont

montré que la quantité :

$$Q(\alpha) = \sum_{x \neq y} B_{xy}\alpha_{xy}$$

étend la modularité de Newman aux classes chevauchantes. La matrice α est définie pour tout système de classes, qu'il s'agisse d'une partition ou d'un recouvrement. Cette formulation permet de mieux comprendre le comportement de cette nouvelle modularité :

- Lorsque la relation α est transitive (cas d'une partition P), $Q(\alpha) = 2m^2 M(P) + 1/2 \sum_{x \in 1..n} d_x^2$. $Q(\alpha)$ est une fonction affine de M, et il revient au même de maximiser M ou Q.
- Lorsque deux classes V_i et V_j sont fusionnées, on modifie les valeurs α_{xy} telles que les éléments $x \in V_i$ et $y \in V_j$ sont nouvellement réunis. Ainsi, on ajoute à la modularité Q la somme des valeurs B_{xy} correspondantes. La modularité croît si et seulement si cette somme est positive.
- Q est bornée supérieurement par la somme des valeurs positives de B : $Q_{max} = \sum_{x \neq y} B_{xy}$. Ainsi Q_{max} est atteint pour tout système de classes constitué des paires (x, y) à valeurs positives dans B, comme les cliques maximales ou les arêtes.

2.2 Une Hiérarchie de Classes Chevauchantes

Dans l'algorithme présenté par Newman pour maximiser la modularité des partitions strictes [6], le point de départ est constitué de l'ensemble des singletons, qui forment une partition de modularité nulle puisqu'il n'y a pas d'arêtes internes. A chaque étape, et tant que la modularité croît, deux classes telles que l'union offre un gain de modularité maximum sont réunies. On fabrique ainsi une hierarchie (arborescence) de classes emboîtées. L'algorithme s'arrête lorsque les classes ne peuvent plus être fusionnées sans faire décroître la modularité.

La formule de modularité Q permet d'étendre ce processus hiérarchique ascendant en partant d'un système de classes chevauchantes et en fusionnant celles-ci : à chaque itération, les classes réunies sont celles qui permettent de maximiser la modularité $Q(\alpha)$ du système de classes résultant. La fusion de deux classes V_i et V_j entraîne leur suppression et l'apparition d'une nouvelle classe $V_i \cup V_j$. Deux systèmes de classes initiales ont d'abord été étudiés :

- Les cliques maximales du graphe : Dans la mesure où elles sont énumérables en un temps raisonnable, elles constituent un système de classes chevauchantes dont la modularité est égale à Q_{max}. Toute fusion fera décroître Q, jusqu'à la valeur $Q_{min} = \sum_{x=1..n} d_x^2$.
- Les arêtes du graphe : On part de la même valeur de modularité Q_{max} et le processus de fusion commence par reconstruire certaines cliques. Dès que l'algorithme ne trouve plus une paire de classes V_i et V_j telle que $\forall (x, y) \in V_i \times V_j, (x, y) \in E$, la modularité commence à décroître.

L'efficacité de l'algorithme ascendant dépend du nombre de classes initiales, puisque celui-ci détermine le nombre d'itérations. En partant des cliques ou des arêtes, un grand nombre d'itérations est effectué, ce qui le rend peu efficace. Un système de classes a alors été défini comme suit :

- Les cliques centrées : En chaque sommet x du graphe, on construit une clique par un algorithme polynomial. On ajoute les sommets potentiels dans l'ordre décroissant des degrés relatifs. Ceci donne une clique contenant x qui est maximale, sans être forcément de cardinal maximum. Au total, après élimination des inclusions, on dispose d'au plus n classes initiales distinctes.

Dans cette procédure ascendante, la modularité varie soit de façon monotone décroissante (cliques maximales, arêtes) soit de façon croissante puis décroissante (cliques centrées). Pour obtenir un ensemble de classes chevauchantes manipulables, le processus de fusion est stoppé pour ne pas se retrouver avec une seule classe, de surcroît de modularité faible. Des critères d'arrêt sont donc introduits, comme celui consistant à borner supérieurement les cardinaux des classes.

2.3 Validation par Simulation

Dans un premier temps, nous avons cherché à valider la méthode sur des graphes artificiels. Nous avons généré des graphes aléatoires à 210 sommets avec 5 classes de 50 éléments. Les 4 premières sont disjointes, et la cinquième est composée de 10 éléments pris dans chacune des 4 premières, plus 10 éléments spécifiques. Dans ces cinq classes, des arêtes sont tirées au hasard avec une probabilité p_i. Il est à noter que lorsque p_i est faible, la simulation s'apparente à une situation de données manquantes. Sur ces graphes, nous lançons l'algorithme hiérarchique, à partir des cliques maximales, des arêtes ou des cliques centrées, jusqu'à obtenir 5 classes, que nous comparons aux classes initales (cf. tableau 1). Pour ces graphes, bien loin d'être des graphes "biologiques", les résultats sont encourageants. On retrouve correctement les classes initiales et les sommets multiples dès que $p_i \geq .20$. Les taux de faux positifs sont bornés à 25%. Enfin, on note que les cliques centrés donnent des classes nettement meilleures pour $p_i = .15$.

	Cliques Maximales				Arêtes				Cliques Centrées			
p_i	$Couv$	$Mult$	$Faux$	Ret	$Couv$	$Mult$	$Faux$	Ret	$Couv$	$Mult$	$Faux$	Ret
.15	.54	29	.22	.57	.49	27	.23	.52	.86	37	.26	.68
.20	.87	46	.24	.87	.70	37	.22	.72	.94	41	.24	.79
.25	.99	53	.25	.99	.87	44	.20	.87	.96	42	.21	.82
.30	1.0	51	.21	1.0	.96	48	.20	.96	.97	41	.17	.85

Tableau 1. Valeurs moyennes des paramètres sur 100 tirages : $Couv$ désigne le taux d'éléments couverts, égal au pourcentage d'éléments retrouvés dans *une* des classes obtenues (couplage de poids maximum) ; $Mult$ le nombre d'éléments classés au moins 2 fois ; $Faux$, le taux de faux positifs, classés plusieurs fois, alors qu'ils ne le sont pas initialement ; Ret, le taux d'éléments multiples retrouvés (à juste titre) classés plusieurs fois.

3 Recherche de Protéines Multifonctionnelles Humaines au sein d'un Interactome Humain

Vingt-cinq PMFs humaines ont préalablement été identifiées par une analyse de la littérature pour constituer un ensemble de test. La méthode ascendante pour classes chevauchantes proposée a été utilisée pour partitionner un graphe d'IPP humaines. Les PMFs de l'ensemble de test ont alors été recherchées dans les classes de la hierarchie de moins de 50 éléments et la pertinence de leur classification a été évaluée par l'analyse des annotations fonctionnelles des classes au sein desquelles elles sont retrouvées.

3.1 Hiérarchie de Classes Chevauchantes sur l'Interactome Humain

Un réseau d'interactions humaines de haute qualité composé de 27276 interactions pour 9596 protéines a été extrait de la base de données APID (bioinfow.dep.usal.es/apid/) grâce à un protocole destiné à éliminer les interactions non physiologiques. Au sein de ce réseau, chaque protéine interagit en moyenne avec $7,8$ autres protéines.

Deux constructions ascendantes ont été calculées à partir de deux systèmes de classes initiales : (i) les cliques maximales, et (ii) les cliques centrées. Le système des arêtes, malgré ses performances intéressantes, a été abandonné car il nécessitait trop d'espace mémoire. Dans les deux cas retenus, le critère d'arrêt est de limiter les classes à 200 protéines afin d'éviter la formation de classes trop importantes pour partager une fonction biologique commune. Le tableau 2 résume les résultats obtenus.

Au sein des deux constructions, le nombre de protéines faisant partie de l'intersection de plusieurs classes finales est grand. En partant de l'ensemble des cliques maximales, les intersections des classes finales contiennent 5082 protéines. Avec les cliques centrées, ce chiffre est nettement inférieur : 2059 protéines.

	Cliques Maximales	Cliques Centrées
Modularité initiale : $Q(\alpha_0)$	1 141 228 679	452 974 653
Nombre de classes initiales	19 408	5 372
Modularité finale : $Q(\alpha_f)$	1 008 635 159	757 685 237
Nombre de classes finales :	109	77

Tableau 2. Evolution de la modularité au cours de la construction de la hiérarchie.

3.2 Annotations Fonctionnelles des Protéines Multifonctionnelles Etudiées

Au sein des deux hiérarchies obtenues, les modules fonctionnels sont mis en évidence. L'algorithme utilisé est identique à celui décrit dans [9]. Brièvement, il s'agit de parcourir la hiérarchie en profondeur en testant, pour chaque sous-arbre, si les protéines formant le sous-arbre partagent une annotation Gene Ontology a telle que les protéines annotées par a soient majoritaires au sein du sous-arbre. Si oui, les protéines de ce sous-arbre constituent alors un "module fonctionnel" annoté par a.
On transfère ensuite aux protéines multiclassées l'ensemble des annotations des modules fonctionnels dans lesquels on les retrouve. Au final, ces annotations peuvent décrire une voire plusieurs fonctions (attention, une protéine multi-annotée n'est pas forcément une PMF). Pour chaque PMF de l'ensemble de test, les annotations/fonctions prédites par la méthode sont alors confrontées à leurs annotations/fonctions connues. Pour cela, plusieurs questions ont été abordées :

1. **Les annotations GO transférées aux PMFs par la méthode couvrent-elles les annotations GO connues pour ces mêmes protéines ?** Quatre-vingt cinq pour cent des annotations GO connues pour les PMFs leurs sont transférées par la méthode dans le cas des cliques maximales contre 60% pour les cliques centrées. Le transfert d'annotations est donc bien efficace.
2. **Les annotations GO transférées aux PMFs correspondent-elles à des fonctions décrites parmi leurs annotations GO connues et ailleurs (en effet, certaines fonctions des PMFs ne sont pas répertoriées sous forme d'annotations GO) ?** Cinquante-huit pour cent et 76% des annotations transférées aux PMFs corroborent les fonctions décrites respectivement dans le cas des cliques maximales et centrées.
   Au vu de ces deux estimateurs, la couverture plus importante à partir du système de cliques maximales (85% vs. 60%) semble réduire le taux de prédictions correctes (58% vs. 76%).
3. **Pour combien de PMFs prédit-on plusieurs fonctions grâce aux annotations transférées ?** Dans le cas des cliques maximales, la multifonctionalité est mise en évidence pour 76% des PMFs contre 64% pour les cliques centrées. La méthode leur prédit plusieurs fonctions différentes décrites dans la littérature. De manière intéressante, on a pu prédire une fonction correcte bien que non décrite dans par leurs annotations GO connues de départ pour 60% d'entre elles.

4 Conclusion et Perspectives

A notre connaissance, une seule autre méthode d'analyse des graphes IPP aboutit à un système de classes chevauchantes : il s'agit de CFinder [10], qui est basée sur une méthode de percolation de cliques [11]. Bien que l'idée paraisse intéressante, la faible densité en arêtes des graphes IPP limite l'efficacité du partitionnement obtenu.

Notre méthode repose sur la construction d'une hiérarchie de classes initiales, chevauchantes. Le choix du critère à optimiser tout comme le choix du système de classes de départ déterminent principalement les résultats obtenus. Afin de pouvoir traiter de grands réseaux tels que l'interactome humain, nous avons proposé et testé un nouveau système de classes de départ, celui des cliques centrées : leur calcul est nettement plus rapide que celui des cliques maximales et le nombre de classes produites est nettement inférieur, ce qui rend la hiérarchie finale plus petite et donc plus facile à exploiter. Nos résultats nous permettent de conclure que bien que la modularité du système de classes finales obtenu à partir des cliques centrées soit inférieure à celle obtenue à partir des cliques maximales, ce système est cependant adapté à la recherche de protéines multifonctionnelles.

Les 25 PMFs étudiées dans ce travail ont toutes été décrites quant à leur multifonctionalité dans de récentes publications. Grâce à cet ensemble de PMFs, nous avons montré que notre méthode identifiait la multifonctionalité de ces protéines dans 76% des cas avec les cliques maximales et 64% des cas avec les cliques centrées. Dans le but de généraliser notre méthode pour identifier des PMFs *ab initio*, nous travaillons actuellement à la mise au point d'un traitement statistique des annotations transférées aux protéines. Cette méthode basée sur la fréquence d'association entre termes GO associés à une même protéine, vise à sélectionner automatiquement des protéines dont les annotations multiples transférées décrivent des fonctions cellulaires distinctes.

Remerciements

CB et AG, agents CNRS, ne remercient pas le Ministère de l'Enseignement Supérieur et de la Recherche pour la politique qu'il tente de mettre en application. Ce travail est financé par le PEPS 2008-2010 du Département ST2I du CNRS.

Références

[1] Jeffery CJ. Moonlighting proteins : old proteins learning new tricks. *Trends Genet.*, 19(8) :415-7, 2003.

[2] Sanchez C, Lachaize C, Janody F, Bellon B, Röder L, Euzenat J, Rechenmann F, Jacq B. Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. *Nucleic Acids Res.*, 27(1) :89-94, 1999.

[3] Aittokallio T, Schwikowski B. Graph-based methods for analysing networks in cell biology. *Brief Bioinform.*, 7(3) :243-55, 2006.

[4] Brucker F, Barthélemy JP. *Eléments de classification : aspects combinatoires et algorithmiques*. Hermès, Paris, 438 p, 2007.

[5] Newman ME. Fast algorithm for detecting community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69 :066133, 2004.

[6] Newman ME. Modularity and community structures in networks. *Proc.Natl.Acad.Sci USA*, 103 :8577-82, 2006.

[7] Angelelli JB, Reboul L. Network modularity optimization by a fusion-fission process and application to protein-protein interactions networks. *Proceedings of JOBIM 2008, Lille, France*, 105-10, 2008.

[8] Brandes U, Delling D, Gaertler M, Gorke R, Hoefer M, Nikoloski Z, Wagner D. On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering*, :20 :172-88, Feb. 2008.

[9] Baudot A, Martin D, Mouren P, Chevenet F, Guénoche A, Jacq B, Brun C. PRODISTIN Web Site : a tool for the functional classification of proteins from interaction networks. *Bioinformatics*, 22(2) :248-50, 2006.

[10] Adamcsek B, Palla G, Farkas IJ, Derenyi I,Vicsek T. CFinder : locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22 :1021-3, 2006.

[11] Palla G, Derenyi I, Farkas I, Vicsek I. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435 :814-8, 2005.

Utilisation d'ontologies de tâches et de domaine pour la composition semi-automatique de services Web bioinformatiques

Nicolas Lebreton[1], Christophe Blanchet[2], Julie Chabalier[1] et Olivier Dameron[1]

[1] InsermU936 - IFR 140, Faculté de Médecine, Université de Rennes 1, 35033 Rennes France
{nicolas.lebreton, olivier.dameron}@univ-rennes1.fr
[2] IBCP UMR 5086, CNRS, Univ. Lyon1, IFR128 BioSciences Lyon-Gerland, 69367 Lyon cedex 07 France
christophe.blanchet@ibcp.fr

Abstract: *Nowadays, bioinformatics tasks typically involve large scale data analysis that require the integration of Web services from heterogeneous platforms. In spite of the efforts for improving Web services interoperability, integration remains difficult and still has to be performed manually by users. Improving the composition of Web services requires to analyze what Web services do as well as the nature and the type of their input and output parameters. This work shows that existing technologies support automating the selection, composition and execution of Web services, and that the current limiting factor to a wider use is the lack of precise enough task and domain ontologies.*

Keywords: Web services, interoperability, task and domain ontologies, OWL-S, SAREK

1 Introduction

L'analyse des données bioinformatiques nécessite de réaliser des enchaînements parfois complexes de traitements. De plus, cela demande éventuellement des étapes intermédiaires de conversion. L'utilisation des services Web pour effectuer chacune de ces opérations facilite l'interopérabilité en bénéficiant de protocoles standards d'échange de données. Ces services Web peuvent être soit autonomes, soit disponibles au sein de grilles de calcul [1] pour des raisons d'efficacité. Dans tous les cas, automatiser une tâche d'analyse de données consiste alors à définir un workflow composant les différents services Web [10]. La création d'un tel workflow repose sur la découverte, la sélection et la composition des services Web. La mise en œuvre de nouveaux workflows comporte alors une part importante d'intervention de l'utilisateur ou des utilisateurs réunissant ces compétences. Pour la composition, il faut faire face à de nombreux obstacles comme les problèmes d'interopérabilité et d'interprétation des rôles des services Web. Toutes les transactions se déroulent à un niveau syntaxique et sont décrites au format WSDL. Cependant, automatiser la définition d'un workflow ou son exécution nécessite également de s'appuyer sur des descriptions sémantiques de ce que font chacun des services ainsi que de la nature de leurs paramètres. Or, il existe actuellement très peu de ces descriptions sémantiques. De nombreux standards existent pour mieux décrire sémantiquement les différents services, notamment OWL-S [1]. Une ontologie de domaine permet de décrire la nature des paramètres des services Web. L'ontologie de formats permet de représenter les formats des données

1. http://www.w3.org/Submission/OWL-S/

du domaine. Une ontologie de tâches [2] permet de décrire la fonction réalisée par les services Web et les conditions nécessaires. Elle fait ainsi appel à une ou plusieurs ontologies de domaine et de formats.

Des projets comme MyGrid [9] et son environnement d'exécution Taverna [4] fournissent des outils pour la conception et l'exécution des workflows grâce à l'utilisation de services Web. La sélection et la combinaison des services sont effectuées manuellement. Seule l'exécution du workflow est automatisée. De plus, il n'y a aucun contrôle sur la cohérence globale du workflow réalisée par l'utilisateur. Taverna s'appuie sur l'ontologie en OWL de MyGrid, qui n'est pas compatible avec OWL-S [9]. De plus, les ontologies de tâches et de domaine de MyGrid ne sont pas assez détaillées pour réaliser automatiquement la sélection et la combinaison des services Web d'un workflow.

2 Objectif

Nous faisons l'hypothèse que les descriptions en OWL-S des services Web bioinformatiques et l'utilisation des technologies associées permettent de dépasser les limitations actuelles de Taverna. L'objectif de ce travail est d'étudier la faisabilité de l'automatisation de la création de workflows classiques en bioinformatiques en utilisant leurs descriptions en OWL-S grâce à une ontologie de tâches et une ontologie de domaine associée.

3 Matériel et méthode

Parmi les 56 services Web disponibles de l'IBCP [2], nous avons selectionné les 16 qui réalisent des recherches de similarité (Blast, Fasta et SSearch) et des alignements multiples (ClustalW). Afin de décrire les services Web, nous complétons leurs descriptions aux formats WSDL indiquant comment communiquer avec ce service par une description en OWL-S. Nous indiquons aussi la tâche réalisée par le service Web ainsi que la nature des paramètres. Pour cela, nous importons OWL-S ainsi que les ontologies de tâches, de domaine et de formats, et nous spécialisons les classes *Service Profile* de OWL-S pour les classes de l'ontologie de tâches.

Pour constituer les ontologies nécessaires à la description sémantique des services, nous allons réutiliser des classes issues de l'ontologie de MyGrid et en créer de nouvelles quand c'est nécessaire. L'ontologie dans MyGrid comporte 475 classes réparties en 6 hiérarchies qui décrivent les tâches, le domaine et les formats. Il est nécessaire de bien filtrer ces différentes informations et de les répartir respectivement dans une ontologie de tâches, de domaine et de formats. Pour l'ontologie de tâches, les classes dans MyGrid étaient trop générales, il était indispensable d'avoir une hiérarchie de classes plus spécifique pour nos services Web. C'est pourquoi, nous avons créé une ontologie de tâches décrivant la tâche effectuée par les différents services Web.

Les différentes ontologies vont nous permettre de définir les services Web et leurs paramètres. L'utilisation d'outil de sélection de services Web comme OWLS-MX [6] montre l'apport des correspondances sémantiques dans la sélection des services Web. OWLS-MX permet la sélection d'un service Web OWL-S spécifique dans un ensemble de services Web. La correspondance entre les différents services Web est basée sur la similarité syntaxique et sémantique des paramètres d'entrées et de sorties d'un service Web. Pour valider la composition semi-automatique des services Web, nous insérons nos services Web sémantiques dans le moteur de composition SAREK [3] qui permet de vérifier l'ordonnancement et l'exécution de l'enchaînement prédéfini. SAREK possède deux modules,

2. http://gbio-pbil.ibcp.fr/ws/

un planificateur qui propose une composition sémantique et un exécuteur qui exécute la composition selon la requête de l'utilisateur. Le planificateur interagit avec l'OPS (Ontology to Publish Services), un répertoire d'ontologies pour découvrir les services Web. L'OPS est décrit en OWL et la description des services Web se fait grâce à OWL-S.

4 Résultats

L'utilisation des ontologies améliore la description des paramètres des 16 services Web et indique leurs tâches respectives. Par exemple, nous allons détailler un échantillon des classes qui sont importantes pour définir les 4 services Web réalisant une recherche de similarité de type Fasta (Tableau 1). Par exemple, nous indiquons que le service *SubmitFasta* réalise une tâche qui consiste à créer un

services Web	Classes pour la tâches	Classes pour le domaine	Classes pour le format
Submit Fasta	creating_processus_ fasta	protein_sequence_database protein_sequence	Fasta_ format
CheckStatusFasta	checking_processus	number_jobid, status_job	
CancelResultsFasta	canceling_processus	number_jobid, aborted_status_job	
GetResults Fasta	web_service_grid fasta (tâche), fasta_grid retrieving_results_ processus_fasta	number_jobid Fasta (algorithme) Fasta_report part_of_the_sequence	

Tableau 1. Description sémantique des 4 service Web de l'IBCP permettant de réaliser un Fasta au moyen de classes provenant des ontologies de tâches, de domaine et de formats.

processus par le biais de la classe *creating_processus_fasta* de l'ontologie de tâches. Les différents paramètres et les formats sont décrits respectivement par les classes de l'ontologie de domaine et de formats. L'apport de ces classes améliore la description syntaxique des paramètres des services. Par exemple, nous pouvons établir que le service *SubmitFasta* prend en entrée une séquence protéique au format fasta et une base de données de séquences protéiques. Le principe de construction reste identique pour les 12 services Web restants (le tableau complet et les services Web en OWL-S sont disponibles en ligne [3]).

Pour la construction de l'ontologie de domaine, il s'agit de récupérer l'ensemble des sous-classes de MyGrid : *bioinformatics_metadata*, *bioinformatics_algorithm*, *bioinformatics_data* et *bioinformatics_data_ressource* qui vont indiquer la nature des entrées et des sorties, les ressources nécessaires et les résultats attendus. Nous filtrons les 413 classes de MyGrid qui se rattachent à l'ontologie de domaine et nous ajoutons 43 classes pour mieux définir le domaine d'application des services Web. Toutes les classes insérées sont situées sous la classe racine *owl*:*thing*. De même, nous récupérons les 33 sous-classes de *bioinformatics_file_formats* de MyGrid et ajoutons 3 classes spécifiques pour représenter l'ontologie de formats. Pour spécifier les éventuelles versions d'un format, il suffit de créer des instances spécifiques des classes décrivant les formats. L'ontologie de tâches permet de trouver les étapes d'une tâche prédéfinie. Nous construisons ainsi une ontologie de tâches de 61 classes dont 25 sont définies et nous ajoutons 93 restrictions et 17 propriétés pour définir les tâches réalisées par les services Web. Par exemple, Les propriétés créées *has_step* et *has_next_step* indiquent les liens entre la tâche composite et les différentes sous-tâches. L'ontologie de tâches spécialise le *service Profile*

3. http://www.ea3888.univ-rennes1.fr/lebreton/SWSDescription.html

d'OWL-S, notamment la classe *profile*:*Profile* (Fig 1). Grâce à l'ontologie de tâches, nous pourrons

Fig. 1. Extrait de l'ontologie de tâches permettant de décrire les services Web.

ainsi exploiter que le fait d'exécuter un Fasta implique de faire une recherche de similarité locale entre des protéines. L'héritage multiple permet de décrire les tâches facilitant ainsi la réutilisation des classes pour d'autres services Web. Au final, l'importation et l'enrichissement des différentes ontologies englobant le domaine, les tâches et les formats totalise 553 classes, 23 propriétés et 300 restrictions (Tableau 2).

Importation et création de classes	Ontologies de tâches	Ontologies de domaine	Ontologies de format
Nombre de classes de MyGrid	0	413	33
Nombre de classes ajoutées	61	43	3
Nombre de classes au total	61	456	36

Tableau 2. Importation et enrichissement des ontologies de tâches, de domaine et de formats.

L'utilisation du logiciel OWLS-MX met en évidence que la recherche d'un service Web est facilitée quand les entrées et les sorties des services Web sont définies sémantiquement par la propriété *process*:*parameterType*. Elle relie les instances représentant les paramètres des services Web aux classes de l'ontologie de domaine. En l'absence de classes décrivant le domaine des services Web nous pouvons uniquement faire de la recherche syntaxique sur le nom du paramètre. L'apport de

la sémantique nous permet d'exploiter automatiquement la hiérarchie des tâches et les éventuelles contraintes associées à chacune afin de faire des recherches plus approfondies et de meilleure qualité. Par exemple, rechercher les services Web réalisant une analyse de séquence (y compris ceux qui ne réalisent qu'un type particulier d'analyse de séquence) dont un des paramètres d'entrée est une séquence protéique.

À partir des descriptions sémantiques des services Web, l'ordre correct des enchaînements des services est retrouvé par le moteur de composition SAREK. Nous choisissons ensuite une tâche dont on connaît au préalable le résultat final. En insérant les différents paramètres lors de l'exécution de l'enchaînement des services, nous retrouvons le résultat attendu par notre analyse.

5 Discussion

Ce travail nous a permis de valider le fait que les techniques actuelles permettent de semi-automatiser la composition des services Web pour réaliser des tâches bioinformatiques classiques. Pourtant, la composition se fait toujours principalement manuellement. Ce travail nous a également permis d'identifier l'absence d'ontologies de tâches et de domaine suffisamment riches comme étant le principal point bloquant à la composition semi-automatique.

La description des services Web demande de définir la nature et le type des entrées et des sorties. En biologie, le format des données joue un rôle important. Nous avons choisi de distinguer deux ontologies, une de domaine et une de formats. Les formats ne sont pas compatibles entre eux, ce qui complique l'interopérablité pour les services Web. La définition des formats n'est pas présente dans la couche syntaxique des services Web (WSDL), il faut donc définir spécifiquement les formats des entrées et des sorties dans la couche sémantique (OWL-S). La sélection d'OWL-S repose sur sa compatibilité avec OWL et le fait qu'elle permet une annotation complète d'un service. Il est à noter que les ontologies crées pour OWL-S sont aussi utilisables par SAWSDL [7] pour l'annotation. En utilisant l'ontologie OWL-S, nous ne pouvons pas définir un héritage multiple au niveau de la propriété *parameterType* du process. La restriction *parameterType exactly one* au niveau de la classe *swrl*:*variable* de OWL-S oblige l'utilisation d'un seul type de paramètre pour définir la classe se référant au domaine. Pour pallier à cette restriction, plusieurs opérations sont possibles. La solution que nous avons choisie consiste à créer une propriété supplémentaire. Pour pouvoir exprimer les formats des paramètres des services, nous utilisons les différentes ontologies de domaine et de formats, mais nous allons réaliser des modifications. Au niveau de l'ontologie de formats, nous avons créé une propriété *has_format_parameter* qui a pour domaine *process*:*parameter* et co-domaine *mygridformat*:*bioinformatics_file_formats*. Pour chaque paramètre, nous pouvons lui spécifier son format en liant les instances ou les classes de l'ontologie de tâches avec la propriété *has_format_parameter*. Cette approche évite la multiplication des classes, des sous-classes et autorise la réutilisation des formats pour d'autres données. Cette solution est idéale pour avoir une ontologie de formats légère et qui ne demande pas une étape de vérification et d'évaluation trop fastidieuse. Il est plus judicieux d'avoir une hiérarchie réutilisable et relativement simple d'utilisation pour l'utilisateur. De plus, il est toujours possible d'utiliser la propriété *process*:*parameterValue* afin qu'elle puisse quand même être exploitée partiellement par d'autres applications compatibles avec OWL-S.

L'apport de l'ontologie de tâches au niveau des services Web permet d'indiquer à l'utilisateur la tâche réalisée par les services Web. De plus, l'identification des services pour accomplir la tâche demandée nécessite de connaître les sous-taches de la tâche globale pour déterminer les différents services utilisées. L'ontologie de tâches permet d'acquérir cette information et donc de décrire la décomposition d'une tâche en sous-tâches.

La composition des services Web est une tâche complexe, elle doit tenir compte de nombreux paramètres comme l'hétérogénéité et la disponibilité des services Web [8]. Il existe différentes approches de composition des services Web. La première approche se base sur les services Web composites qui définissent un ensemble de services atomiques et la façon dont ils communiquent entre eux. Dans la deuxième approche, la composition est vue comme de la planification en générant automatiquent un plan d'exécution des services Web. Nous avons utilisé SAREK pour l'évaluation de la combinaison des services Web, mais il existe d'autres logiciels pour la planification comme OWLS-XPLAN, SHOP2 [5]. SAREK est le premier cadre de travail qui fournit une tolérance pour les pannes de services dans la composition des services. En effet, quand un service subit un échec, SAREK peut proposer un autre service. Une autre composition peut-être choisie et exécutée, si un service manque. En biologie, il existe de nombreux services Web réalisant la même tâche, mais ils sont localisés de manière hétérogène. Dans ce cadre, il serait intéressant de classifier plusieurs services Web réalisant la même tâche pour permettre le remplacement d'un service en cas d'échec par un service équivalent.

Nous avons démontré l'apport des ontologies de tâches, de domaine et de formats dans la sélection et la composition semi-automatique des services Web bioinformatiques. Ces ontologies sont réutilisables pour la réalisation d'autres scénarios biologiques, à la différence de scripts écrits à la main. Il serait judicieux d'enrichir ces modélisations en ajoutant d'autres services Web afin d'obtenir des enchaînements permettant de résoudre des problèmes biologiques plus complexes.

6 Remerciement

Ce travail est co-financé par la commission européenne par le biais du projet EMBRACE, dans le cadre de la thématique "Life sciences, genomics and biotechnology for health". Le numéro du contrat au sein du 6ème programme cadre de l'UE est le LHSG-CT-2004-512092.

Références

[1] C. Blanchet, C. Combet, G. Deleage, Integrating Bioinformatics Resources on the EGEE Grid Platform, *Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid*, IEEE Computer Society, pp. 48, 2006.

[2] B. Chandrasekaran, J.R. Josephson, R. Benjamins, The Ontology of Tasks and Methods, *Proceedings of the 11th Knowledge Acquisition Modeling and Management Workshop*, Banff, Canada, 1998.

[3] D.B. Claro, R.J.A. Macedo, Dependable Web Service Compositions using a Semantic Replication Scheme, *Proceedings of the XXVI Brazilian Symposium of Networks and Distributed Systems, SBRC*, Rio de Janeiro, 1 :441-454, 2008.

[4] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, et al, Taverna : a tool for building and running workflows of services, *Nucleic Acids Research*, 34 :W729-32, 2006.

[5] M. Klusch, Semantic Web Service Coordination, *CASCOM - Intelligent Service Coordination in the Semantic Web*. Birkhaeuser Verlag, Springer, 2008.

[6] M. Klusch, B. Fries, Hybrid OWL-S Service Retrieval with OWLS-MX : Benefits and Pitfalls, *Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web at ISWC*, 2007.

[7] J. Kopecky, T. Vitvar, C. Bournez, J. Farrell, SAWSDL : Semantic Annotations for WSDL and XML Schema, *IEEE Internet Computing*, 11 :60-67, 2007.

[8] R. Stevens, C.A. Goble, S. Bechhofer, Ontology-based knowledge representation for bioinformatics, *Briefings in Bioinformatics*, 1 :398-414, 2000.

[9] K. Wolstencroft, P. Alper, D. Hull, C. Wroe, et al., The myGrid ontology : bioinformatics service discovery, *International Journal of Bioinformatics Research and Applications*, 3 :303-325, 2007.

[10] C. Wroe, C. Goble, A. Goderis, P. Lord, et al., Recycling workflows and services through discovery and reuse : Research Articles, *Concurrency and Computation : Practice & Experience*, 19 :81-194, 2007.

Probabilistic modeling of tiling array expression data

Aurélie Leduc[1], Stéphane Robin[2], Philippe Bessières[1] and Pierre Nicolas[1]

[1] INRA, Mathématique Informatique et Génome UR1077
F-78350 Jouy-en-Josas, France
pierre.nicolas@jouy.inra.fr, aurelie.leduc@jouy.inra.fr,
philippe.bessieres@jouy.inra.fr
[2] AgroParisTech/INRA, Mathématiques et Informatique Appliquées UMR 518
16 rue Claude Bernard, F-75005 Paris, France
robin@agroparistech.fr

Abstract: *For organisms with small genomes such as bacteria, the current microarray technology allows adopting a tiling design where the whole genome is covered by overlapping probes. These arrays permit to measure the transcriptional activity of the whole genome with unprecedented resolution. Model-based approaches currently used to analyze these data remain however very simple, the most popular model being the piecewise constant Gaussian model with a fixed number of breakpoints. Here we present a new approach based on hidden Markov modelling designed for the probabilistic reconstruction of the trajectory of a continuous-valued signal. The use of this model does not require the choice of a fixed number of breakpoints and permits to account for subtle effects such as drift in the signal. The model also includes direct correction for the variations of probe affinities via the use of covariates.*

Keywords: Expression data, tiling arrays, hidden Markov models.

1 Introduction

The tiling design for oligonucleotide microarrays consists of overlapping probes that provide uniform covering of the genomic sequence. Their hybridization with RNA samples (cDNA), allow to assess the transcriptional activity of the whole genome of organisms such as bacteria and yeasts with high resolution [2,10]. The continuous improvement of the technology renders these arrays more affordable. Generalization of the use of such arrays should greatly improve our understanding of the complexity and the dynamics of transcriptional landscapes. This context justifies the improvement of the currently available statistical methods dedicated to the analysis of tiling array transcription data.

The problem of the analysis of these data is naturally stated in terms of finding segments where the hybridization signal is relatively constant, delimited by breakpoints that are expected to correspond to biological features such as promoters, terminators or splicing sites. A variety of tools including local non-parametric smoothing [11,14] and simple iterative hypothesis testing [7] have been proposed to answer this question. Today the most popular and best mathematically grounded model is the piecewise constant model with Gaussian noise [8,4]. The simplicity of this approach is appealing but its use presents a number of specific difficulties, the two most obvious being the choice of the number of segments and the high time complexity of the algorithm. Partial answers for each of these two problems are found in [8] and [4], respectively.

In principle, embedding the segmentation model in a probabilistic setting that includes not only the noise but also the evolution of the signal can alleviate the need for the choice of a fixed number of breakpoints. In this context the problem states as the estimation of a parameter and the reconstruction of the underlying signal trajectory can integrate the uncertainty on the exact number of breakpoints. This idea stimulated the development of Hidden Markov models (HMMs) [3,5,6,13]. However, transcript level is a continuous quantity and none of the proposed models is satisfactory when the signal of interest is continuous. A HMM achieving this aim at a computationally affordable cost will be presented here.

The proposed model is also markedly richer than the piecewise constant model. First, it automatically accounts for differential affinity between probes via the introduction of covariates. This allows to achieve segmentation and within-array normalization in one step. Second, our model also relaxes the assumption of strictly constant underlying signal between abrupt "shifts" by also allowing progressive "drift".

2 Methods

2.1 Model

Like in previous approaches [7,8,4], x_t, the $\log_2$ of the observed intensity at position t, is modeled as the sum of an unobservable signal u_t that is the focus of interest plus a Gaussian noise with standard deviation σ. This general model can be written

$$x_t \mid u_t \sim \mathcal{N}(u_t, \sigma^2)\,. \tag{1}$$

However, u_t is not seen in our model as a parameter but is itself a random variable. Correlation between probes that are adjacent on the chromosome is accounted for by a Markov transition kernel $\pi(u_t, u_{t+1})$ and $(x_t, u_t)_{1\leq t\leq n}$ is thus said to be a hidden Markov model [9,1]. Compared with traditional use of HMMs, the complication comes from the continuous nature of u_t whereas the efficient algorithmic machinery of the HMMs (Viterbi algorithm, forward-backward algorithm, EM algorithm) works well for discrete and typically small number of hidden states [9]. In general, with K hidden states, the time complexity of the algorithms is $O(nK^2)$.

Here we propose a structure of the transition matrix $\pi(u_t, u_{t+1})$ accounting for abrupt shifts and progressive drifts in the unobservable signal u_t that allows to discretize the continuous range $U_{\min} \leq u_t \leq U_{\max}$ in K points spaced by a regular interval, $h = (U_{\max} - U_{min})/(K-1)$. This particular structure warrants time complexity $O(nK)$ for the classical HMM algorithms and thus permits appropriately high resolution of discretization.

For values of u_t and u_{t+1} taken on the internal points of the discretized hidden state space, the transition probability writes

$$\begin{aligned}\pi(u_t, u_{t+1}) = {} & \alpha_n I_{\{u_{t+1}=u_t\}} + \alpha_s \eta_h(u_{t+1}) \\ & + \alpha_u I_{\{u_{t+1}>u_t\}} \lambda_u^{\frac{u_{t+1}-u_t}{h}-1}(1-\lambda_u) \\ & + \alpha_d I_{\{u_{t+1}<u_t\}} \lambda_d^{\frac{u_t-u_{t+1}}{h}-1}(1-\lambda_d)\,,\end{aligned} \tag{2}$$

where the parameters verify $0 \leq \alpha_n, \alpha_s, \alpha_u, \alpha_d \leq 1$, $\alpha_n + \alpha_s + \alpha_u + \alpha_d = 1$ and $0 \leq \lambda_u, \lambda_d < 1$ and with $I_{\{X\}}$ standing for 1 if X is true, 0 otherwise. This transition kernel is best understood as a mixture of four types of moves with weights α_n, α_s, α_u and α_d. The parameter α_n accounts for

unchanged u between successive probes. Shift moves have probability α_s and the distribution of the signal after the move is independent of the value of the signal before the move. This distribution is given by η_h and it approximates the marginal distribution of the signal. Namely, $\eta_h(u_{t+1}) = \int_{u_{t+1}-h/2}^{u_{t+1}+h/2} \eta(u)du$, where η is the kernel density estimate computed on x with a Gaussian kernel and Scott's bandwidth [12]. The possibility of small drift, either upward or downward, is accounted for by α_u and α_d. Drift amplitudes are modeled by two geometric distributions of parameters λ_u and λ_d and average amplitudes write $h + h/(1-\lambda)$.

It can be verified that as $h \to 0$ and $h/(1-\lambda) \to \gamma$ the transition kernel of the discrete-valued Markov chain of Equation 2 converges in distribution towards the transition kernel of a continuous-valued Markov chain. With an appropriately high K it should thus be possible to approach, using the discrete-valued model of Equation 2, the results that one would obtain with the continuous-valued model.

Genomic DNA hybridization data was used in a pre-processing step by Huber *et al.* (2006) for the purpose of between-probe signal normalization and outlier trimming. The model proposed here accounts for these effects by modeling the genomic DNA hybridization intensities as a covariate. The probability distribution for the observed variable x_t given the underlying signal u_t and the gDNA residuals r_t writes as a mixture model

$$\begin{aligned} x_t \mid u_t, r_t \sim (1-\epsilon(r_t))\mathcal{N}(u_t + \rho(u_t)r_t, \sigma(u_t)^2) \\ +\epsilon(r_t)\mathcal{U}(U_{\min}, U_{\max})\,, \end{aligned} \tag{3}$$

where $\epsilon(r_t)$ corresponds to the probability of outliers, $\mathcal{U}(U_{\min}, U_{\max})$ is the uniform distribution that models outlier data and $\mathcal{N}(u_t + \rho(u_t)r_t, \sigma(u_t)^2)$ is the Gaussian distribution modeling non-outlier data. This model is markedly richer than Equation 1. Notice (i) the non-constant proportionality factor $\rho(u_t)$ applied to r_t, (ii) the non-constant standard error $\sigma(u_t)$ of the Gaussian distribution, (iii) the probability of outliers ϵ that depends on r_t. More precisely, ρ and σ are modeled as piecewise constant function of u_t with 8 intervals, and ϵ is a two-parameter logistic function of the absolute value of r_t, $\epsilon(r_t) = 1/(1+e^{-(a+b|r_t|)})$.

2.2 Algorithms

The particular structure of transition matrix defined by 2 allow $O(nK)$ implementations of the HMM classical algorithms, namely

1. likelihood computation ($P(x_{1\ldots n})$),
2. forward-backward algorithm (computation of $P(u_t|x_{1\ldots n})$ for each t),
3. Viterbi algorithm (finding the trajectory $u_{1\ldots n}$ that maximizes $P(u_{1\ldots n}|x_{1\ldots n})$).

These algorithms are implemented in our software. All the parameters are estimated in the Maximum Likelihood (ML) framework with the EM algorithm, an iterative algorithm that alternates an E-step (forward-backward algorithm) and a M-step (parameter update). The output provides a detailed report on the "denoised" signal based on the results of the Viterbi and forward-backward algorithms.

3 Example of application

3.1 Data set

Our example data-set used here comes from pilot experiments conducted on *Bacillus subtilis* within the European consortium BaSysBio [10]. This array consists of 383 149 probes starting every 22 nt

on each strand of the *B. subtilis* genome (GenBank: AL009126). Probe lengths range between 45 nt and 65 nt and were adjusted to reduce TM variations (isothermal design). Production of the tiling arrays, synthesis of labeled cDNA from the RNA samples with random priming, hybridization and signal acquisition were carried out by Nimblegen. RNA was extracted from *B. subtilis* culture during exponential growth on rich medium. One out of four biological replicates gave a high quality signal and is analyzed here [10].

3.2 Discretization and parameter estimates

The model was designed with the explicit aim of modeling a continuous-valued underlying signal. In other words, discretization of the hidden state space is seen only as a necessary technicality and the step $h \propto 1/K$ should ideally be sufficiently small to have no impact on the results. Intuitively, the smaller the standard-deviation of the noise σ, the smaller the step h should be. The results obtained on the *B. subtilis* data-set confirm this intuition and thereby provide some form of validation for the model (the data will be shown in the oral presentation).

Parameter estimates in model-based analyzes are an invaluable source of information to better understand both the behavior of the model and the data. The shape of the transition matrix that describes the trajectory of the underlying signal is defined by the parameters in Equation 2. We found a high value of α_n: it is estimated that the underlying signal remains unchanged between adjacent probes in more than 85% of the cases. The upward and downward drift moves accounted for most of the remaining cases. The value of the parameters (α_u) and (α_d), corresponded to 5.0% and 7.8%, respectively. The proportion of abrupt shift was estimated to be much smaller (1.2%).

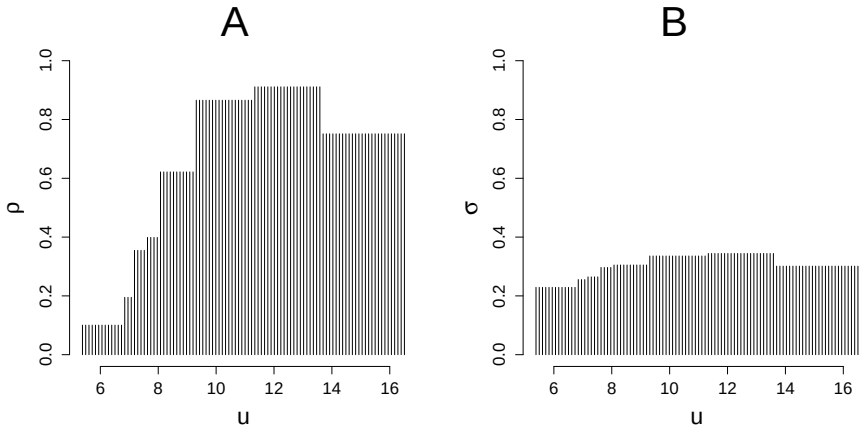


Figure 1. Parameter estimates. A: proportionality factor ρ applied to r_t as a function of the signal level u_t. B: Standard-deviation of the noise σ as a function of the underlying signal level u_t.

Examination of the parameter values also revealed the importance of modeling ρ as a function of of the underlying signal level u_t (a eight-parameter piecewise constant function). The relationship between ρ and u_t is represented in Figure 1A. By contrast, the standard deviation of the noise σ is a relatively flat function of u_t as shown in Figure 1B.

These results emphasize the importance of two specificities of our model: the modeling of drift moves as a complement to shift moves and the non-constant ρ that provides a simple adaptive method to account for the variation of affinity between probes.

3.3 Reconstruction of the "denoised" signal

The adoption of a probabilistic setting for the trajectory of the underlying signal allows for a considerably richer signal reconstruction than just "optimal" trajectory reconstruction. Figure 2 gives an illustration of these possibilities by superimposing a number of results obtained with the model on a $10\,000$ bp region of the *B. subtilis* chromosome. Results include: (i) the prediction interval for the value of the signal u_t at each chromosome position; (ii) a point prediction for the signal value by the conditional mean of u_t (the best predictor in terms of quadratic error); (iii) the inferred position of the experimental point after correction for differential probe affinity (computed as $x_t - \hat{\rho}(\hat{u}_t)r_t$); (iv) the exact position of each type of move in the best trajectory given by the Viterbi path (abrupt shift, upward drift and downward drift); (v) the probability of having each type of move at each position.

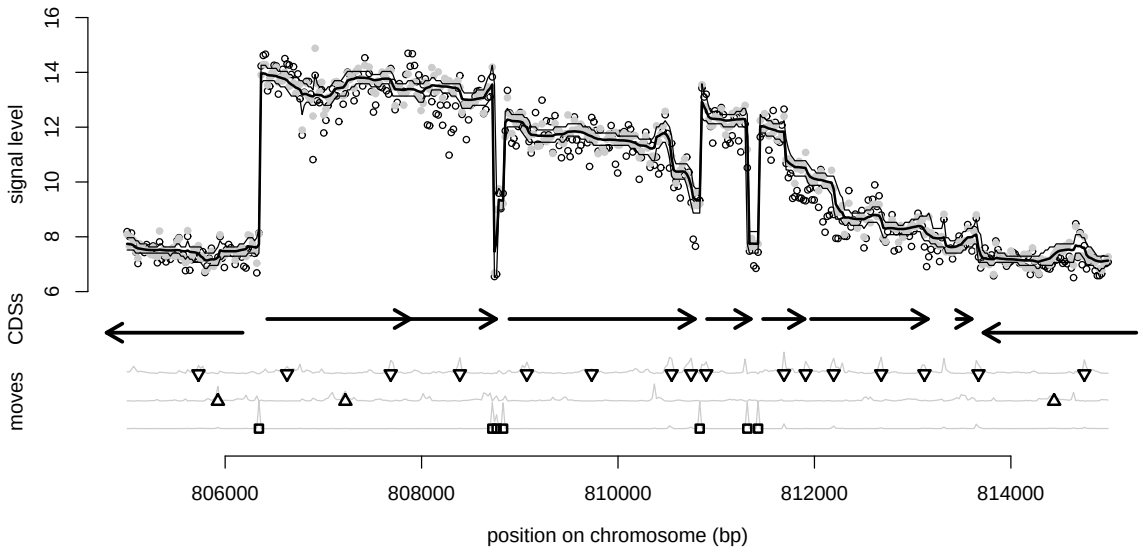


Figure 2. Reconstruction of the transcriptional landscape. One strand of a $10\,000$ bp segment of the *B. subtilis* chromosome is represented. Upper part: Open circles show the original signal. Closed gray circles represent the signal after "correction" with the genomic DNA covariate. The thick black line shows the expectation of the transcript level as computed with the HMM. Thin black lines correspond to the 95% CI. Middle part: Horizontal arrows indicate GenBank CDSs. Lower part: Shift moves along the most likely trajectory are shown as squares. Upward and downward drift moves are indicated by point-up and point-down triangles, respectively. Move probabilities are represented as gray lines.

The biological pertinence of the distinction between shifts and drifts seems remarkable in Figure 2. Inferred shifts are found mostly in intergenic regions that *a priori* correspond to possible positions for transcriptional promoters and terminators. The meaning of the drift is, on the contrary, not obvious. Drift might partly reflect local variations of labeled cDNA that result from technical artifact such as random priming bias. The asymmetry with more downward than upward drift may not be unexpected in this case. Interestingly, drift could also reflect biological differences in the

amount of mRNA. Asymmetry may then, for instance, be caused by molecules whose synthesis is still incomplete.

The algorithm has also been successfully tested on tiling array data obtained with Affymetrix technology [2].

4 Conclusion

We describe a new methodology based on a hidden Markov model that embeds the segmentation of a continuous-valued signal in a probabilistic setting. For a computationally affordable cost, this framework alleviates the difficulty of choosing a fixed number of breakpoints and permits retrieving more information than a unique segmentation. Probabilistic modeling makes it straightforward to compute confidence measures on the estimated transcriptional landscape. This information should prove particularly useful to pinpoint the differences in large collections of arrays. This will be discussed in the oral presentation.

Acknowledgments

We are grateful to Hanne Jarmer and Simon Rasmussen for providing us the data-set. This work is supported by the BaSysBio project, European Commission research grant (LSHG-CT2006-037469).

References

[1] Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) Biological sequence analysis. Cambridge University Press.

[2] David,L., Huber,W., Granovskaia,M., Toedling,J., Palm,C.J., Bofkin,L., Jones,T., Davis,R.W. and Steinmetz,L.M. (2006) A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. USA*, **103**, 5320–5325.

[3] Fridlyand,J., Snijders,A.M., Pinkel,D., Albertson,D.G., and Jain,A.N. (2004). Hidden Markov Model Analysis of Array CGH Data. *J. Multivariate Analysis* **90**, 132-153.

[4] Huber,W., Toedling,J., and Steinmetz,L.M. (2006), Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, **e22**, 1963-1970.

[5] Marioni,J.C., Thorne,N.P. and Tavaré,S. (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.

[6] Munch,K., Gardner,P.P., Arctander,P., and Krogh,A. (2006) A hidden Markov model approach for determining expression from genomic tiling micro arrays. *BMC Bioinformatics*, **7**, e239.

[7] Olshen,A.B., Venkatraman,E.S., Lucito,R., Wigler,M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557-572.

[8] Picard,F., Robin,S., Lavielle,M., Vaisse,C., and Daudin J.-J. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, e27.

[9] Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257-286.

[10] Rasmussen,S., Nielsen,H.B. and Jarmer,H. (submitted) Transcriptionally active regions in the genome of *Bacillus subtilis*. Submitted.

[11] Royce,T.E., Carriero,N.J. and Gerstein,M.B. (2007) An efficient pseudomedian filter for tiling microrrays. *BMC Bioinformatics*, **8**, e186.

[12] Scott,D.W. (1992) Multivariate Density Estimation. Theory, Practice and Visualization. New York: Wiley.

[13] Stjernqvist,S., Rydén,T., Sköld,M., and Staaf,J. (2007) Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*, **23**, 1006-1014.

[14] Wang,L.-Y., Abyzov,A., Korbel,J.,O., Snyder,M., and Gerstein,M.B. (2009) MSB: A mean-shift-based approach for the analysis of structural variation in the genome. *Genome Research*, **19**, 106-117.

Master regulator analysis reveals key transcription factors for Germinal Center formation

Celine Lefebvre, Mariano J. Alvarez, Presha Rajbhandari, Wei Keat Lim and Andrea Califano

Center for Computational Biology and Bioinformatics, Columbia University,
1130 St Nicholas ave, New York NY 10032 USA
{lefebvre, califano}@c2b2.columbia.edu

Abstract: *We describe a new method for the identification of master regulators of a phenotype of interest. The master regulator analysis identifies transcription factors that are candidate master regulators of a phenotype of interest based on its transcriptional targets. We applied this method for deciphering the regulation of Germinal Center B cell programs, revealing the two transcription factors MYB and FOXM1 as synergistic master regulators.*

Keywords: Interactome, Master Regulator, Germinal Center.

1 Introduction

The identification of transcription factors driving the transition from a phenotype to another is a challenge that have traditionally been addressed by looking at their differential expression between the two biological conditions. Here we propose a method for the identification of transcription factors (TFs) that are master regulators of a phenotype of interest by using the TF's targets, or regulon, as a proxy for its activity, by looking at its enrichment in differenitally expressed genes in the phenotype of interest. We applied this method to the discovery of master regulators of Germinal Center B cells.

Upon T-dependent antigen activation, Germinal Center (GC) B cells undergo somatic hypermutation of their immunoglobulin genes and selection of cells that have acquired increased affinity for the antigen. When compared to their naïve precursors, genes involved in cell proliferation, DNA metabolism and apoptosis (pro-apoptotic program) are over-expressed in GC B cells (i.e., GC-activated), while others, including anti-apoptotic genes, cytokines/chemokines, cell adhesion-related genes, and inhibitors of cell proliferation are downregulated in GC B cells (GC-repressed) [1]. While the transcriptional repressor BCL6 was shown to be necessary for GC formation [2], its activity is not sufficient to justify the complex genetic program changes in the GC and the TFs that choreograph these changes are still largely undetermined. For instance, it is unclear how the hyper-proliferative phenotype observed in the GC is achieved in the absence of the C-MYC protein [1], a key regulator of cellular metabolism and growth.

We have previously developed a Human B Cell Interactome (HBCI), which represents ~66,000 known and predicted protein-protein (PPIs) and protein-DNA interactions (PDIs) using a Bayesian evidence integration approach [3,4]. The HBCI provides a systems level representation of the transcriptional machinery supporting GC formation and maintenance. Here, we used the

transcriptional network of the HBCI to discover master regulators of GC. Among others, the MYB and FOXM1 TFs were shown to play a key role in regulating GC cell-cycle related programs.

2 Method

2.1 Human B Cell Interactome

We have assembled a multi-layer Human B Cell Interactome (HBCI), representing approximately 66,000 mature B cell-specific transcriptional, signaling, and protein-complex interactions, using an established Bayesian Evidence Integration Approach [3,5]. The HBCI constitutes a unique resource for a Human cell context and integrates evidence supporting specific interactions from multiple, heterogeneous sources, both computationally inferred and experimental. These include, among others, a large collection of 254 B cell Gene Expression Profiles representative of normal and tumor related mature B cell phenotypes [3], protein-protein interaction from experimental assays and databases, literature datamining, and inferences from reverse engineering algorithms, such as ARACNe [6-9] and MINDy [4,10,11]. Each evidence source is assigned a prior likelihood, proportional to the probability of observing the specific evidence source in a large set of Gold Standard Positive interactions that are experimentally validated. Likelihoods for distinct evidence sources are then combined, using the Bayes theorem, to produce a single posterior likelihood representing the probability that a specific interaction is a true positive.

2.2 Master Regulator Analysis

The master regulator analysis, or MRA, is based on two components: (1) a transcriptional interaction network and (2) a gene expression profile with samples of two phenotypes A and B, for example a tumor type and the corresponding normal cell type. MRA attempts to identify TFs inducing the transition from A to B. Each TF in the transcriptional network is associated with a set of targets that can be either activated or repressed by the TF. Therefore, we defined a positive (R_{TF}^{+}) and a negative (R_{TF}^{-}) regulon per TF, respectively defined as the set of TF-activated and TF-repressed targets and computed using the Spearman correlation between the TF and its targets in the gene expression profile. Specifically, the R_{TF}^{+} and R_{TF}^{-} targets of an activated TF in phenotype B should be respectively up- and down-regulated in B when compared to A (the opposite for a repressed TF). The advantage of regulon analysis is that it is independent of the type of TF activation (i.e. transcriptional, post-transcriptional, or post-translational). Indeed, it is independent of the TF mRNA expression. Clearly, one does not expect all targets in a regulon to be equally informative: some targets may lack key co-factors, necessary for expression in a specific cellular context; others may be false positives and thus not representative of the TF's activity. Thus, to infer activated TFs, one can analyze the statistical enrichment of the R_{TF}^{+} and R_{TF}^{-} regulons in genes that are respectively up- and down-regulated in B compared to A (or vice-versa if testing for repressed TFs). This can be accomplished using Gene Set Enrichment Analysis algorithm (GSEA) [12].

2.3 Gene Set Enrichment Analysis

GSEA uses the Kolmogorov-Smirnov statistical test to assess whether a predefined gene set is

statistically enriched in genes that are at the two extremes of a list ranked by differential expression between two biological states that we call a reference list [12]. The algorithm is very useful to detect differential expression of a set of genes as a whole, even though the fold-change may be small for each individual gene. The reference list is ranked with the T-statistics computed by comparing phenotypes B and A. For each TF in the transcriptional network, GSEA is applied independently to the R_{TF}^{+} and R_{TF}^{-} to compute their enrichment and the best enrichment p-value is associated to the TF. GSEA null distribution is computed by permuting gene labels in the reference list 10,000 times.

This enrichment analysis is effective in detecting TFs with significant activity change but ineffective in ranking them, because regulon size dramatically affects enrichment p-values. We first selected enriched TFs, or Master Regulators (MRs) based on the p-value and then classified the MRs by their Differentially Expressed Target Odds Ratio (DETOR score), defined as:

$$DETOR_{MR_i} = \frac{\left(GS_i^{LE} / RS_i^{LE}\right)}{\left(GS_i / RS\right)},$$

where GS_i^{LE} and RS_i^{LE} are the number of genes in the gene set and the reference set before the leading edge defined by GSEA, and GS_i and RS are the sizes of the gene set and the reference set. This score provides a direct assessment of the percent of regulon genes that are differentially expressed in B compared to A and thus of the TF regulon specificity to the phenotype B. Here we use the DETOR score to rank regulons that have been previously selected as significantly enriched in differentially expressed genes by their GSEA p-values, therefore the DETOR score is higher than 1. DETOR score is computed for R_{TF}^{+} and R_{TF}^{-} and the best DETOR score is used for MR classification.

2.4 Shadow and Synergy

Two interesting cases arise when MRs have significant overlap in their regulon, and we introduce here the notion of Shadow and Synergistic MRs. A Shadow Regulator (SR) is inferred as statistically significant only because its regulon overlaps that of a *bona fide* MR. In this case, the SR is no longer significant if the MR targets are ignored. However, two MRs may have a significant regulon overlap and be involved in synergistic regulation. In this case, their common targets will be more enriched than those of the individual TFs. To test the SR effect, we first consider the complete list of significant MRs ranked by DETOR score, remove the regulon of the best MR from all other regulons and retest the new regulons for enrichment with GSEA. Based on the new p-value associated to each MR, the enrichment is either still significant and the MR is kept for the next iteration or not significant anymore and the MR is considered as shadowed by the best MR. We iterate this process until all MRs have been considered. At each step, we create a new cluster that is identified by the candidate MR (best cluster DETOR score) and includes all candidate SRs shadowed by the MR.

Once the clusters have been identified, we can test MR-pairs for synergism, including both best-MR/SR pairs in a cluster (to identify MRs that would have otherwise been discarded as SRs), and MR-MR pairs in different clusters. Based on the analysis, individual and synergistic MRs are candidate regulators controlling formation of phenotype B, while SRs can either be MR-controlled TFs involved in feed-forward loops with the MRs or *bona fide* artifacts of regulon overlap.

3 Results

3.1 Master Regulators of Germinal Center Reaction

The HBCI and the MRA are based on a B cell gene expression profile representative of 18 normal and neoplastic human B cell phenotypes [3]. We considered 101 TFs from the HBCI that are associated with regulons of at least 100 targets. The positive (R_{TF}^{+}) and the negative (R_{TF}^{-}) regulons, respectively defined as the set of TF-activated and TF-repressed targets, were defined using the Spearman correlation between the TF and its targets in the complete B cell gene expression profile. The reference list for GSEA was created by comparing Centroblast samples to Naïve samples with a t-test. GSEA enrichment ($p < 0.01$ after Bonferroni correction) identified 24 GC-activated MRs (a-MRs) and 47 GC-repressed MRs (r-MRs). By DETOR ranking, the most significant GC-activated MR is MYB (short for C-MYB), DETOR = 5.43, I.e., more than 58% of its regulon genes were found to be differentially expressed in the GC. Among the a-MRs, MYB, NFYB, E2F5, HMGA1 and E2F1 had high DETOR score (>4), while among the r-MRs, IRF9, FOXJ2, RXRA, IRF5, and NR1H2 had similar rank. This analysis recapitulates most TFs previously reported to play a role in GC formation, including BCL6 [2] and LMO2 [13] among a-MRs, and P53 [14], IRF4 [15] and POU2F2 [16] among r-MRs. While results between differential activation and differential expression of the MR are generally in agreement, ranking of top a-MRs and r-MRs was substantially different. A puzzling result was the identification of MYC as an a-MR, despite our previous report of MYC expression loss in the GC [1]. This suggests that other TFs may contribute to the regulation of MYC targets in GC. The Shadow analysis identified 42 clusters, where each cluster is identified by the candidate MR with best cluster DETOR score and includes all candidate SRs shadowed by the MR. a-MRs were tested for synergistic activity, including both MR/SR pairs in each cluster and MR/MR pairs in distinct clusters. The analysis led to the identification of 4 synergistic TF-pairs, regulating more than 100 targets. Each synergistic pair included FOXM1, respectively paired with MYB, NFYB, E2F5 and E2F1 (ranked by DETOR score), the MYB/FOXM1 was the most significant synergistic a-MR pair (by DETOR score). More specifically, the MYB/FOXM1 regulon includes 150 common targets, compared to 222 MYB-targets and 1,287 FOXM1-targets. 133 of 150 common targets were differentially expressed in the GC ($FDR \leq 0.05$ by Mann-Whitney test), showing almost complete specificity (89%) of their combinatorial program. We thus proceeded to experimentally test the GC-specific activity of the MYB/FOXM1 pair, using both quantitative Chromatin Immunoprecipitation (qChIP) assays and lentivirus mediated shRNA silencing assays.

3.2 MYB is a Transcriptional Activator of FoxM1

To assess whether MYB and FOXM1 may transcriptionally regulate each other, we silenced each TF independently in ST486 Burkitt's Lymphoma cells and monitored their mRNA and protein levels over a 72 hour time course. ST486 cells transduced with lentiviral particles encoding FOXM1, MYB, and control shRNA were harvested at 24h, 48h and 72h post-infection. Decrease of endogenous MYB and FOXM1 protein levels, compared to cells transduced with control shRNA, was confirmed by Western blots. Additionally, qRT-PCR assays showed that both TFs were effectively silenced at the mRNA level after 24 hours and continued to be silenced at 48h and 72h. While MYB mRNA and protein levels were not affected by FOXM1 silencing, FOXM1 mRNA and protein levels were reduced in a time-dependent manner following MYB silencing, suggesting that MYB is a transcriptional activator of FOXM1. We further confirmed that MYB binds to the FOXM1 promoter by qChIP assays.

3.3 MYB and FOXM1 Common Targets Validation

We tested whether the MYB/FOXM1 regulon was affected by the silencing of either TF in isolation. Gene expression profile from the Affymetrix HG-U95A GeneChip® was obtained from ST486 cell lines at 24 hours after transduction with lentiviral particles carrying FOXM1, MYB, and control shRNA in three replicates each. Silencing was confirmed by Western blot and qRT-PCR. The 24h time point was selected to ensure that FOXM1 protein level was not yet affected by MYB silencing (confirmed by Western blot), thus ruling out indirect regulation via FOXM1. Gene expression profile data from MYB and FOXM1 silenced cells, as well as negative controls, were normalized with GCRMA (GC Robust Multi-array Average) [17]. Differentially expressed genes were then ranked by t-test analysis. Enrichment of the predicted common targets in the HBCI against the differentially expressed genes was then assessed by GSEA. Experiments confirmed the dramatic enrichment of the 126 genes in the positive MYB/FOXM1 regulon among downregulated genes, upon silencing of either TF ($p \leq 10^{-4}$ in both cases).

To further characterize the regulation of the target genes, we performed qChIP on four predicted common targets in the HBCI for which the mRNA levels was decreasing significantly after FOXM1 or MYB silencing experiments. The results showed that both FOXM1 and MYB bind to the promoter of these genes to regulate their transcription.

4 Conclusion

Our validation suggests that regulon analysis is a substantially better predictor of TF activity than TF mRNA level. Indeed regulon-based analysis is independent of whether the TF-activity was modulated at the transcriptional or post-transcriptional/translational level. Additionally, moderate differential expression of a TF, may lead to significant differential expression of its target genes. Thus the proposed analysis has the advantage of being independent of the expression profile of the TF, depending only on the transcriptional activity of the TF's targets.

Taken together, these data show that the HBCI and the Master Regulator analysis can be a useful tool in the elucidation of important physiologic phenotypes, such as the germinal center reaction. In follow up work, we are also investigating its ability to dissect pathways that are dysregulated in disease.

Acknowledgements

This work was supported by the National Cancer Institute (R01CA109755), the National Institute of Allergy and Infectious Diseases (R01AI066116), and the National Centers for Biomedical Computing NIH Roadmap Initiative (U54CA121852).

References

[1] U. Klein, Y. Tu, G.A. Stolovitzky, J.L. Keller, J. Haddad, Jr., V. Miljkovic, G. Cattoretti, A. Califano and R. Dalla-Favera Transcriptional analysis of the B cell germinal center reaction, Proc Natl Acad Sci U S A 100 (2003) 2639-2644.

[2] T. Fukuda, T. Yoshida, S. Okada, M. Hatano, T. Miki, K. Ishibashi, S. Okabe, H. Koseki, S. Hirosawa, M. Taniguchi, N. Miyasaka and T. Tokuhisa Disruption of the Bcl6 gene results in an impaired germinal

center formation, J Exp Med 186 (1997) 439-448.
[3] C. Lefebvre, W.K. Lim, K. Basso, R. Dalla-Favera and A. Califano A context-specific network of protein-DNA and protein-protein interactions reveals new regulatory motifs in human B cells, Lecture Notes in Bioinformatics 4532 (2007) 42-56.
[4] K.M. Mani, C. Lefebvre, K. Wang, W.K. Lim, K. Basso, R. Dalla-Favera and A. Califano A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas, Mol Syst Biol 4 (2008) 169.
[5] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt and M. Gerstein A Bayesian networks approach for predicting protein-protein interactions from genomic data, Science 302 (2003) 449-453.
[6] K. Basso, A.A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera and A. Califano Reverse engineering of regulatory networks in human B cells, Nat Genet 37 (2005) 382-390.
[7] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera and A. Califano ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, BMC Bioinformatics 7 Suppl 1 (2006) S7.
[8] T. Palomero, W.K. Lim, D.T. Odom, M.L. Sulis, P.J. Real, A. Margolin, K.C. Barnes, J. O'Neil, D. Neuberg, A.P. Weng, J.C. Aster, F. Sigaux, J. Soulier, A.T. Look, R.A. Young, A. Califano and A.A. Ferrando NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth, Proc Natl Acad Sci U S A 103 (2006) 18261-18266.
[9] A.A. Margolin, K. Wang, W.K. Lim, M. Kustagi, I. Nemenman and A. Califano Reverse engineering cellular networks, Nat Protoc 1 (2006) 662-671.
[10] K. Wang, N. Banerjee, A.A. Margolin, I. Nemenman and A. Califano Genome-wide discovery of modulators of transcriptional interactions in human B lymphocytes, Lecture Notes in Computer Science 3909 (2006) 348-362.
[11] K. Wang, M. Alvarez, B. Bisikirska, R. Linding, K. Basso, R. Dalla Favera and A. Califano Dissecting the Interface Between Signaling and Transcriptional Regulation in Human B Cells, Pac Symp Biocomput 14 (2009).
[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, Proc Natl Acad Sci U S A 102 (2005) 15545-15550.
[13] Y. Natkunam, S. Zhao, D.Y. Mason, J. Chen, B. Taidi, M. Jones, A.S. Hammer, S. Hamilton Dutoit, I.S. Lossos and R. Levy The oncoprotein LMO2 is expressed in normal germinal-center B cells and in human B-cell lymphomas, Blood 109 (2007) 1636-1642.
[14] R.T. Phan and R. Dalla-Favera The BCL6 proto-oncogene suppresses p53 expression in germinal-centre B cells, Nature 432 (2004) 635-639.
[15] G. Cattoretti, R. Shaknovich, P.M. Smith, H.M. Jack, V.V. Murty and B. Alobeid Stages of germinal center transit are defined by B cell transcription factor coexpression and relative abundance, J Immunol 177 (2006) 6930-6939.
[16] K. Schubart, S. Massa, D. Schubart, L.M. Corcoran, A.G. Rolink and P. Matthias B cell development and immunoglobulin gene transcription in the absence of Oct-2 and OBF-1, Nat Immunol 2 (2001) 69-74.
[17] W.K. Lim, K. Wang, C. Lefebvre and A. Califano Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks, Bioinformatics 23 (2007) i282-288.

Cellular automata modeling of intercellular genetic regulatory networks

Anne Crumière

Institut de Mathématiques de Luminy (UMR 6206)
Campus de Luminy, Case 907 13288 MARSEILLE Cedex 9
crumiere@iml.univ-mrs.fr

Abstract: *Biologists often represent genetic interactions by directed graph, named interaction graph. Vertices represent genes, whereas edges represent regulatory effects from one gene on another. Edges are labelled with a positive sign in the case of an activation and negative for an inhibition. This article deals with relationships between the structure of such graphs and their dynamical properties.*
The biologist R.Thomas enounced, thirty years ago, the following two general rules: a necessary condition for multistability is the presence of a positive circuit in the interaction graph (the sign of a circuit being the product of the signs of its edges) and the existence of a negative circuit is a necessary condition for the presence of sustained oscillations. These rules are about the dynamic of a single cell, and it has given rise to mathematical statements and proofs. This article aims at extending these rules to regulatory interactions spanning within cells and between cells in the discrete formalism.

Keywords: Spatial differentiation, intercellular genetic regulatory network, interaction graph, discrete formalism, multistability, cellular automata.

1 Introduction

Les biologistes représentent souvent les réseaux de régulation génétique par des graphes. Ces graphes, appelés graphes d'interactions, sont des graphes orientés et signés, notés $G = (V, E)$, où l'ensemble des nœuds, $V = \{1, \ldots, n\}$, représente les n gènes du système et l'ensemble des arêtes E représente les régulations : $(i, j) \in E$ si le gène i est un régulateur du gène j. Les arêtes sont signées, positivement $(+1)$ dans le cas d'une activation, c'est à dire quand la protéine codée par le gène i favorise l'expression du gène j, et négativement dans le cas d'une inhibition (-1), c'est à dire quand la protéine codée par le gène i ralentit ou stoppe l'expression du gène j. Ce papier traite des relations entre la sructure de tels graphes et leurs propriétés dynamiques.

On s'intéresse principalement à des propriétés de stabilité de ce système ; dans le cadre des systèmes dynamiques, cela se traduit par les notions d'attracteurs, d'états stationnaires, de cycles attractifs... L'étude de ces propriétés demande soit dans un cadre continu la résolution d'un système d'équations différentielles assez grand, soit dans le cadre discret, l'étude d'un graphe de grande taille représentant la dynamique. Dans les deux cas, on est confronté à des problèmes de très grande complexité.

D'où l'idée de revenir aux graphes d'interactions qui correspondent à ces dynamiques, et qui ont l'avantage d'être de taille plus petite et donc plus facile à analyser. De tels liens entre la dynamique et

les graphes d'interactions existent : les ***règles de Thomas*** en sont un bel exemple. En effet, dans les années 80, le biologiste René Thomas énonça deux règles [8] :

- une condition nécessaire pour l'existence de plusieurs états stationnaires dans la dynamique est la présence d'un circuit positif dans le graphe d'interactions (un ***circuit*** dans ce graphe étant une séquence de sommets $(i_1, \dots, i_r) \in V$ telle que $(i_k, i_{k+1}) \in E$ pour $k \in \{1, \dots, r\}$ avec $i_{r+1} = i_1$ par convention et le ***signe d'un circuit*** étant le produit des signes des arêtes),
- une condition nécessaire pour la présence d'oscillations stables ou amorties dans la dynamique est la présence d'un circuit négatif dans le graphe d'interactions.

Que ce soient la multistabilité ou les oscillations, ils correspondent tous deux à d'importants phénomènes biologiques : processus de différenciation cellulaire et homéostasie respectivement.

Durant cette décennie, de nombreux chercheurs ont formalisé et démontré ces règles dans des cadres différents [4,5,6], mais toujours dans le cas où les gènes sont répartis dans une même cellule. Ce travail étend dans un cadre discret ces règles à un réseau génétique intercellulaire. Dans un premier article [2], on a consideré un ruban infini de cellules avec une communication intercellulaire locale (de la cellule centrale vers ses voisines gauche-droite). Cette communication locale, qui est biologiquement raisonnable, est standard et à la base des automates cellulaires. Dans cet article, nous présentons une version plus générale de ce cadre : les cellules sont réparties sur un réseau de dimension quelconque et la communication intercellulaire est étendue à un voisinage quelconque. Dans ce cadre, nous montrons les deux règles de Thomas, avec des hypothèses spatiales supplémentaires pour le cas de la règle positive. La règle négative proposée ici dans un cadre intercellulaire mais valide aussi dans un cadre intracellulaire est un affinement du théorème proposé dans [7]. Les preuves et plus de détails sont disponibles dans [3,1]

2 Formalisme

2.1 Dynamique du réseau de régulation

On s'intéresse à l'évolution d'un système composé de cellules, chaque cellule contient la même collection de gènes choisis dans un ensemble fini I. Pour un gène $i \in I$, l'intervalle $\mathcal{A}_i = [0, k_i]$ désigne les niveaux d'expression possibles du gène i. Un ***état*** d'une cellule est un élément du produit cartésien $\mathcal{A} = \prod_{i \in I} \mathcal{A}_i$.

Généralement, un système biologique est constitué de plusieurs cellules. On peut supposer que les cellules sont réparties régulièrement et disposées suivant un réseau $\mathbb{M}$, i.e. un sous-groupe discret de $\mathbb{R}^d$ muni de l'opération $+$. Chaque cellule est dans un état $a \in \mathcal{A}$. Un ***état du système*** est ainsi une suite d'éléments de $\mathcal{A}$ indexée par $\mathbb{M}$, i.e. un élément de $\mathcal{A}^{\mathbb{M}}$. Pour tous $s \in \mathcal{A}^{\mathbb{M}}$ et $\mathbb{U} \subset \mathbb{M}$, on note $s_{\mathbb{U}}$ la ***restriction*** de s à $\mathbb{U}$. Lorsque l'état du système est donné par $s \in \mathcal{A}^{\mathbb{M}}$, l'état de la cellule $x \in \mathbb{M}$ est noté $s(x)$ et pour un gène $i \in I$, le niveau d'expression du gène i dans la cellule x est noté $s(x, i)$.

Le niveau d'expression d'un gène dans une cellule varie au cours du temps en fonction des niveaux d'expression des gènes dans cette cellule et dans les cellules voisines. Chaque cellule communique avec ses voisines de manière uniforme dans l'espace. Pour modéliser ce phénomène, on considère l'ensemble fini $\mathbb{V} \subset \mathbb{M}$ appelé ***voisinage*** et une ***fonction locale*** $f : \mathcal{A}^{\mathbb{V}} \to \mathcal{A}$. On suppose que l'élément nul de $\mathbb{M}$ appartient à $\mathbb{V}$. La ***dynamique globale*** du système peut alors être donnée par l' ***automate cellulaire*** $F : \mathcal{A}^{\mathbb{M}} \to \mathcal{A}^{\mathbb{M}}$ défini par $F(s)(x) = f((s(x + v))_{v \in \mathbb{V}})$ pour tous $s \in \mathcal{A}^{\mathbb{M}}$ et $x \in \mathbb{M}$. On désigne par $F(s)(x, i)$ la valeur vers laquelle le niveau d'expression du gène i dans la cellule x tend quand le système est à l'état s. Tout naturellement, un ***état stationnaire*** de F est un état

$s \in \mathcal{A}^{\mathbb{M}}$ tel que $F(s) = s$, *i.e.* un état dans lequel le niveau d'expression de chaque gène n'évolue pas.

EXEMPLE 2.1. *(Réseau hexagonal)*
On étudie un système constitué de deux gènes par cellule, i.e, $I = \{a, b\}$, *qui ont deux niveaux d'expression, ainsi* $\mathcal{A} = \{0,1\} \times \{0,1\}$. *Les cellules sont localisées sur un plan. Pour des raisons biologiques, les cellules sont représentées par des hexagones qui pavent le plan. Ce* ***pavage hexagonal*** $\mathbb{M}$ *est généré par les vecteurs* $\mathbf{e_1}$ *et* $\mathbf{e_2}$, *plus précisement* $\mathbb{M} = \mathbb{Z}\mathbf{e_1} + \mathbb{Z}\mathbf{e_2}$. *Soit* $(x_1, x_2) \in \mathbb{Z}^2$, (x_1, x_2) *sont les coordonnées de la cellule* $x \in \mathbb{M}$ *dans le pavage selon une origine arbitraire* O *et la base* $(\mathbf{e_1}, \mathbf{e_2})$ *(voire Figure 1). Le voisinage d'une cellule quelconque* x *de coordonnées* (x_1, x_2) *est donc :*

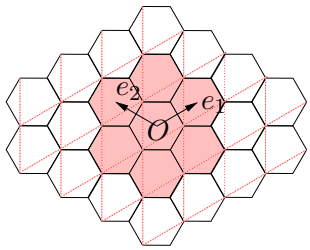


Fig. 1: *Localisation*

$x+\mathbb{V} = \{(x_1, x_2), (x_1, x_2+1), (x_1+1, x_2+1), (x_1-1, x_2), (x_1+1, x_2), (x_1-1, x_2-1), (x_1, x_2-1)\}$.

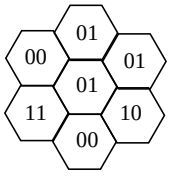


Fig. 2: $x_{\mathbb{V}}$

L'état local $s_{\mathbb{V}}$, *representé en Figure 2, est composé de 7 cellules hexagonales où les deux nombres dans chaque cellule sont les niveaux d'expression des deux gènes contenus dans chaque cellule. Cet état est mathématiquement représenté par la matrice suivante :* $\begin{pmatrix} & (0,0) & (0,1) \\ (1,1) & (0,1) & (0,1) \\ (0,0) & (1,0) & \end{pmatrix}$. *Plus généralement pour tout état* $s \in \mathcal{A}^{\mathbb{M}}$, *l'état* $s(x + \mathbb{V})$ *est une matrice* 3×3 *avec deux trous :*

$$s(x+\mathbb{V}) = \begin{pmatrix} & s(x_1, x_2+1) & s(x_1+1, x_2+1) \\ s(x_1-1, x_2) & s(x_1, x_2) & s(x_1+1, x_2) \\ s(x_1-1, x_2-1) & s(x_1, x_2-1) & \end{pmatrix}$$

Par la suite, pour des raisons d'espace, nous supposerons qu'il y a un seul gène dans chaque cellule, i.e, $I = \{a\}$ *avec toujours deux niveaux d'expression,* i.e. $\mathcal{A}_a = \{0,1\}$. *Une partie de la dynamique locale de ce nouveau système est donnée en Figure 3.*

$f\begin{pmatrix} 0&0 \\ 0&0&0 \\ 0&0 \end{pmatrix} = 0$	$f\begin{pmatrix} 0&0 \\ 0&0&1 \\ 0&0 \end{pmatrix} = 0$	$f\begin{pmatrix} 0&0 \\ 0&1&0 \\ 0&0 \end{pmatrix} = 1$	$f\begin{pmatrix} 0&0 \\ 0&0&1 \\ 0&1 \end{pmatrix} = 0$	$f\begin{pmatrix} 0&0 \\ 0&1&0 \\ 1&0 \end{pmatrix} = 1$	$f\begin{pmatrix} 0&0 \\ 0&0&1 \\ 1&1 \end{pmatrix} = 0$
$f\begin{pmatrix} 0&0 \\ 0&1&0 \\ 1&1 \end{pmatrix} = 0$	$f\begin{pmatrix} 0&0 \\ 1&0&0 \\ 1&1 \end{pmatrix} = 0$	$f\begin{pmatrix} 0&0 \\ 0&1&1 \\ 1&1 \end{pmatrix} = 1$	$f\begin{pmatrix} 0&0 \\ 1&1&1 \\ 0&1 \end{pmatrix} = 1$	$f\begin{pmatrix} 0&1 \\ 0&0&0 \\ 0&1 \end{pmatrix} = 0$	$f\begin{pmatrix} 0&1 \\ 0&0&0 \\ 1&1 \end{pmatrix} = 1$
$f\begin{pmatrix} 0&1 \\ 0&1&0 \\ 0&1 \end{pmatrix} = 0$	$f\begin{pmatrix} 0&1 \\ 0&1&0 \\ 1&0 \end{pmatrix} = 1$	$f\begin{pmatrix} 0&1 \\ 0&0&1 \\ 1&1 \end{pmatrix} = 1$	$f\begin{pmatrix} 0&1 \\ 0&1&0 \\ 1&1 \end{pmatrix} = 1$	$f\begin{pmatrix} 0&1 \\ 1&1&0 \\ 1&0 \end{pmatrix} = 1$	$f\begin{pmatrix} 0&1 \\ 0&1&1 \\ 0&1 \end{pmatrix} = 1$
$f\begin{pmatrix} 0&1 \\ 0&1&1 \\ 1&0 \end{pmatrix} = 1$	$f\begin{pmatrix} 0&1 \\ 0&1&1 \\ 1&1 \end{pmatrix} = 1$	$f\begin{pmatrix} 0&1 \\ 1&1&0 \\ 1&1 \end{pmatrix} = 1$	$f\begin{pmatrix} 0&1 \\ 1&1&1 \\ 1&1 \end{pmatrix} = 0$	$f\begin{pmatrix} 1&0 \\ 0&0&0 \\ 1&0 \end{pmatrix} = 0$	$f\begin{pmatrix} 1&0 \\ 1&1&0 \\ 1&0 \end{pmatrix} = 1$
$f\begin{pmatrix} 1&1 \\ 1&1&0 \\ 0&0 \end{pmatrix} = 1$	$f\begin{pmatrix} 1&1 \\ 1&0&0 \\ 1&0 \end{pmatrix} = 0$	$f\begin{pmatrix} 1&1 \\ 0&1&0 \\ 1&1 \end{pmatrix} = 1$	$f\begin{pmatrix} 1&1 \\ 0&1&1 \\ 1&1 \end{pmatrix} = 1$		

Fig. 3. Dynamique locale

2.2 Dynamique asynchrone

Deux choix de dynamiques sont possibles pour la "mise à jour" du système : une dynamique synchrone (tous les gènes évoluent simultanément) ou une dynamique asynchrone (seul un gène va mettre à jour son niveau d'expression). L'hypothèse synchrone n'est pas biologiquement recevable : l'augmentation ou la diminution du niveau d'expression d'un gène demande un certain délai et l'hypothèse synchrone impose que tous ces délais soient identiques, ce qui est peu probable. En particulier, aucune différence est faite entre le processus de régulation intracellulaire d'un côté, et d'un autre côté la régulation due à la diffusion. C'est pourquoi René Thomas [8] décrit dans le cadre intracellulaire une dynamique asynchrone à partir de la fonction globale. Dans notre cas, on peut décrire la dynamique asynchrone à partir de la dynamique globale donnée par F.

- Soit $t \in \mathbb{Z}$, on définit $\text{sg}(t) = 0$ si $t = 0$, $\text{sg}(t) = +1$ si $t > 0$ et $\text{sg}(t) = -1$ si $t < 0$.
- Soient $(s, s') \in (\mathcal{A}^{\mathbb{M}})^2$ et $(x, i) \in \mathbb{M} \times I$, on définit $s^{(x,i) \lhd s'}$ par : pour tout $(y, j) \in \mathbb{M} \times I$,

$$s^{(x,i) \lhd s'}(y, j) = \begin{cases} s(y, j) & \text{si } (x, i) \neq (y, j), \\ s(y, j) + \text{sg}(s'(y, j) - s(y, j)) & \text{sinon.} \end{cases}$$

Soit un automate cellulaire $F : \mathcal{A}^{\mathbb{M}} \to \mathcal{A}^{\mathbb{M}}$, la ***dynamique asynchrone non-déterministe*** est un graphe, nommé graphe de transition asynchrone, $GTA(F)$, défini par :

- l'ensemble des sommets est $\mathcal{A}^{\mathbb{M}}$, chaque sommet représente un état possible du système,
- il y a une arête de s vers s', quand il existe une cellule $x \in \mathbb{M}$ et un gène $i \in I$ tels que $F(s)(x, i) \neq s(x, i)$ et $s' = s^{(x,i) \lhd F(s)}$.

Ce système évolue donc d'un état $s \in \mathcal{A}^{\mathbb{M}}$ vers un autre état $s' \in \mathcal{A}^{\mathbb{M}}$ suivant les arêtes de $GTA(F)$: le niveau d'expression d'au plus un gène, en au plus une cellule, est changé à chaque pas.

REMARQUE 2.2. *La principale propriété dynamique auquelle nous nous intéressons ici, est la présence d'états stationnaires, qui est indépendante du choix de la dynamique : synchrone, asynchrone...*

2.3 Graphe d'interactions

Généralement, on observe les variations du niveau d'expression d'un gène lorsque les autres gènes interagissent avec celui-ci. Ces variations sont mises en évidence par le calcul de la Jacobienne discrète et visualisées sous la forme d'un graphe d'interactions.

DÉFINITION 2.3. *Soient $s, s' \in \mathcal{A}^{\mathbb{M}}$ et $((x, i), (y, j)) \in (\mathbb{M} \times I)^2$, on définit la* ***Jacobienne discrète*** *de F en $s \in \mathcal{A}^{\mathbb{M}}$ suivant la direction $s' \in \mathcal{A}^{\mathbb{M}}$ comme étant la matrice dont les coefficients sont :*

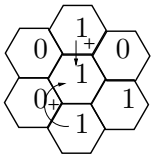
$$\partial_{(x,i),(y,j)} F(s, s') = sg(s'(x, i) - s(x, i)) sg(F(s^{(x,i) \lhd s'})(y, j) - F(s)(y, j)).$$

Pour visualiser ou plutôt représenter les différentes actions d'un gène sur un autre dans une même cellule ou dans une cellule voisine dans une région $\mathbb{U} \subset \mathbb{M}$, on définit le ***graphe d'interactions*** de $\partial F(s_{\mathbb{U}}, s'_{\mathbb{U}})$, noté $G(\partial F(s_{\mathbb{U}}, s'_{\mathbb{U}}))$. Le graphe d'interactions est un graphe orienté signé, *i.e*, avec un signe $+1$ ou -1, attaché à chaque arête et défini par :

- les sommets sont les gènes de chaque cellule contenue dans $\mathbb{U}$, c'est à dire $\mathbb{U} \times I$,
- il y a une arête allant du gène i dans la cellule x vers le gène j dans la cellule y si $\partial_{(x,i),(y,j)} F(s_{\mathbb{U}}, s'_{\mathbb{U}}) \neq 0$ et $\text{sg}(F(s)(y, j) - s(y, j)) \neq \text{sg}(F(s^{(x,i) \lhd s'})(y, j) - s(y, j))$.
  Le signe de l'arête est alors celui de $\partial_{(x,i),(y,j)} F(s_{\mathbb{U}}, s'_{\mathbb{U}})$, il est positif (resp. négatif) si le niveau d'expression du gène i est en "augmentation" (resp. "diminution").

EXEMPLE 2.4. *(Réseau hexagonal) Le graphe d'interactions suivant est construit par calcul des dérivées :*

$$G\left(\partial F\left(\begin{pmatrix}0&1\\0&1&0\\1&1\end{pmatrix},\begin{pmatrix}1&0\\1&0&1\\0&0\end{pmatrix}\right)\right) =$$



3 Conditions nécessaires de multistabilité

DÉFINITION 3.1. *On rappelle que $\mathbb{V}$ est le voisinage de la fonction globale F. Soient $\mathbb{U}', \mathbb{U}''$ deux sous-ensembles de $\mathbb{M}$, on définit le sous-ensemble $\mathbb{U}' + \mathbb{U}'' = \{u' + u'' : u' \in \mathbb{U}' \text{ et } u'' \in \mathbb{U}''\}$. Pour $\mathbb{U} \subset \mathbb{M}$ fini, on note $\partial\mathbb{U} = (\mathbb{U} + \mathbb{V}) \setminus \mathbb{U}$ le* ***bord*** *de $\mathbb{U}$.*

Notre formalisme pour décrire les interactions génétiques dans un cadre intercellulaire est maintenant mis en place. Nous pouvons énoncer l'adaptation de la règle de Thomas au système intercellulaire avec la même collection de gènes dans chaque cellule :

THÉORÈME 3.2. *[1] Soient $F : \mathcal{A}^{\mathbb{M}} \to \mathcal{A}^{\mathbb{M}}$ une fonction globale et $\mathbb{U} \subset \mathbb{M}$. Si $r, t \in \mathcal{A}^{\mathbb{M}}$ sont deux états stationnaires qui vérifient $(r)_{\partial\mathbb{U}} = (t)_{\partial\mathbb{U}}$ et $(r)_{\mathbb{U}} \neq (t)_{\mathbb{U}}$, alors il existe $s \in \mathcal{A}^{\mathbb{M}}$ tel que $G(\partial F(s,t))$ a un circuit positif élémentaire.*

REMARQUE 3.3. *Comme $G(\partial F(s,t))$ contient un circuit élémentaire positif mais comme $s_{\partial\mathbb{U}} = t_{\partial\mathbb{U}}$, le circuit positif est localisé sur $\mathbb{U}$.*

EXEMPLE 3.4. *(Réseau hexagonal) F a deux états stationnaires r et t répondant aux conditions du Théorème 3.2 (une partie uniquement de chaque état stationnaire est représentée). L'ensemble $\mathbb{U}$ est constitué des deux cellules non colorées entourées par les cellules colorées et $\partial\mathbb{U}$ est constitué des huit cellules colorées. Donc il existe un état s tel que $G(\partial F(s,t))$ a un circuit positif élémentaire.*

$r =$ $t =$ $G\left(\partial F\left(\begin{pmatrix}0&1\\0&\mathbf{1}&0\\0&\mathbf{1}&1\\1&1\end{pmatrix},\begin{pmatrix}0&1\\0&\mathbf{0}&0\\0&\mathbf{0}&1\\1&1\end{pmatrix}\right)\right) =$

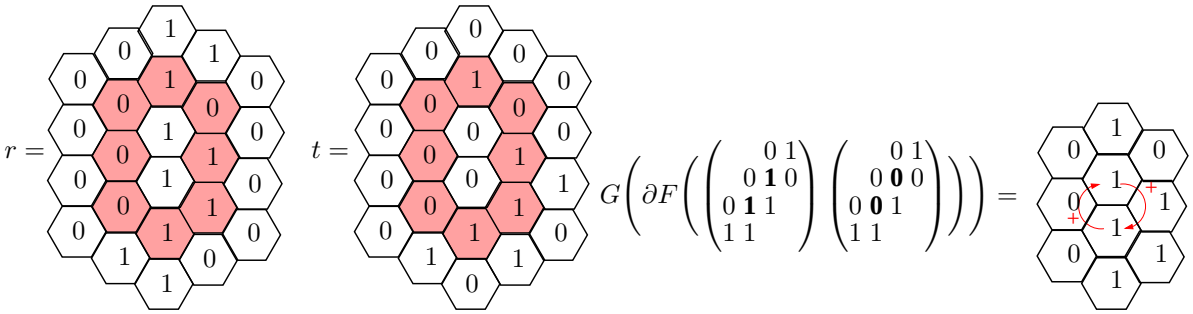


Fig. 4. (Gauche) Etats stationnaires (Droite) Circuit positif entre les gènes des deux cellules centrales

4 Conditions nécessaires pour la présence d'oscillations stables

Dans le cas intracellulaire, la présence d'oscillations stables implique la présence d'un circuit négatif dans le graphe d'interactions [5,7]. On montre dans cette section que l'hypothèse d'oscillations stables dans le graphe asynchrone est très restrictive. En effet, l'existence d'un certain chemin dans le graphe asynchrone suffit à impliquer la présence d'un circuit négatif dans l'union des graphes d'interactions associés à un certain nombre d'état.

THÉORÈME 4.1. *[1] Soient $F : \mathcal{A}^{\mathbb{M}} \to \mathcal{A}^{\mathbb{M}}$ et un chemin $(s^0, s^1, \ldots, s^r)$ dans $GTA(F)$. On définit l'ensemble C tel que :*

$C = \{(p,(x,i),(y,j)) \in \{0,\ldots,r-2\} \times (\mathbb{Z}\times I) \times (\mathbb{Z}\times I)$ *avec* $(x,i) \neq (y,j)$ *tel que* $s^p(x,i) \neq s^{p+1}(x,i)$ *et* $s^{p+1}(y,j) \neq s^{p+2}(y,j)\}$.

Si le chemin vérifie les deux conditions suivantes :

- *il existe* $(x,i) \in \mathbb{Z} \times I$ *tel que* $s^r(x,i) \neq s^{r-1}(x,i)$ *et* $s^1(x,i) \neq s^0(x,i)$ *et* $sg(s^r(x,i) - s^{r-1}(x,i)) \neq sg(s^1(x,i) - s^0(x,i))$
- *pour tout* $(p,(x,i),(y,j)) \in C$, *on a* $sg(F(s^p)(y,j) - s^p(y,j)) \neq sg(F(s^{p+1})(y,j) - s^{p+1}(y,j))$

alors $\bigcup_{(p,(x,i),(y,j))\in C} G(\partial F(s^p, s^{p+1})$ *contient un circuit négatif.*

Regardons de plus près les conditions requises sur le chemin. La première demande que le chemin débute et termine suivant la même direction (le même gène varie), mais dans des sens opposés (les variations doivent être de signe différent). Pour la deuxième condition, il faut d'abord remarquer que l'ensemble C désigne tous les changements de direction du chemin (les différents gènes qui varient au cours du parcours du chemin). La deuxième condition demande qu'à chaque changement de direction du chemin, il soit impossible de tourner dans le graphe asynchrone juste avant dans la même direction et dans le même sens que le changement de direction en cours.

REMARQUE 4.2. *Ce théorème et sa démonstration s'adaptent naturellement au cas intracellulaire, il suffit d'omettre le numéro des cellules. Ce théorème apporte ainsi une précision supplémentaire sur la présence des circuits négatifs dans les graphes d'interactions par rapport aux théorèmes existants (qui restent eux-même valides dans ce cadre intercellulaire).*

5 Perspectives

Cet article propose un modèle discret d'un réseau génétique intercellulaire et donne une adaptation des deux règles de Thomas dans ce nouveau cadre. Dans le Théorème 3.2, la localisation du circuit est induite par les états stationnaires, donc dépendante de la dynamique du système. Mais dans la plupart des modèles étudiés par les biologistes, les circuits sont localisés sur au plus deux cellules. Il serait intéressant d'étudier le nombre maximun de cellules sur lesquel le circuit positif s'étend.

De plus, ce modèle positionne les cellules sur un réseau, donc régulièrement dans l'espace. Une perspective serait de les répartir sur une structure moins ordonnée mais suffisament régulière pour définir une dynamique locale, par exemple le pavage de Penrose.

Références

[1] A. Crumière, Circuits de rétroaction dans les réseaux génétiques de régulation intercellulaires. *PhD Thesis*, 2008.

[2] A. Crumière and P. Ruet, Spatial differentiation and positive circuits in a discrete framework.*ENTCS Series*,92 : 85-100, 2008.

[3] A. Crumière and M. Sablik, Positive circuits and multidimensional spatial differentiation : Application to the formation of sense organs in Drosophila. *BioSystems*, Volume 94, Issues 1-2, October-November 2008, Pages 102-108.

[4] C. Soulé, Graphic requirements for multistationarity. *ComPlexUs*, 1 :123–133, 2003.

[5] É. Remy, P. Ruet, and D. Thieffry, Graphic requirements for multistability and attractive cycles in a Boolean dynamical framework. *Advances in Applied Mathematics* 41(3) : 335-350, 2008.

[6] A. Richard, Modèle formel pour les réseaux de régulation génétique et influence des circuits de rétroaction. *PhD Thesis*, 2006.

[7] A. Richard, On the link between oscillations and negative circuits in discrete genetic regulatory networks. *Proceedings de JOBIM*, 213-218, Marseille 2007.

[8] R. Thomas, On the relation between the logical structure of systems and their ability to generate multiple steady states and sustained oscillations. *Series in Synergetics*, 9 :180-193. Springer, 1981.

Using Reliable and Surprising Item Sets for the Characterization of Protein-Protein Interfaces

Christine Martin[12], Antoine Cornuéjols[2]

[1] LIMSI-CNRS, Université Paris-Sud, UPR CNRS 3251
Bâtiments 508 et 502bis 91403 ORSAY (France)
christine.martin@limsi.fr
[2] AgroParisTech, Equipe Statistique et génome, UMR AgroParisTech/INRA 518
Département M.M.I.P., 16 rue Claude Bernard 75005 Paris Cedex France
antoine.cornuejols@agroparistech.fr

Abstract: *Numerous research effort have been aimed to characterize and predict protein-protein interfaces. This paper introduces a method using only known protein-protein interfaces and combining frequent item set mining techniques with statistical tests to ensure the selection of interesting features. Starting from a database of known interfaces described with geometrical elements, the method produces the elements and combinations thereof that are* characteristic *of the interfaces. This approach allows one to eliminate the need for negative instances and to come up with easy to interpret features, as compared to techniques that operate as "black-boxes". The results obtained on a set of 459 protein-protein interfaces from the DOCKGROUND database confirm that the findings are consistent with current knowledge about protein-protein interfaces.*

Keywords: Protein-protein docking, data mining, frequent itemsets.

1 Introduction

Being able to predict protein-protein interactions is of paramount importance for biology and medicine [6,10]. A first approach is to *start from fundamental premisses*, i.e. the sequences of amino-acids that make up the proteins, and knowledge of their physico-chemical properties and how these translate in terms of energy. In principle, a sufficiently detailed model should allow one to compute with enough accuracy the energy of each configuration of interest and therefore predict the likely protein-protein complexes and their probable binding sites. This line of attack is however precluded, at least at the present time, by the sheer magnitude of the size of the search space and by the complexity of the energy computations. Another route is to *learn by automatic means to discriminate positive protein-protein complexes from negative ones* [3]. One issue is the choice of representation of the protein-protein complexes. Another one stems from the fact that, usually, databases only contain positive instances. Negative instances have to be generated, often using random conformations and orientations of the proteins, assuming that these correspond to bad or impossible pairings. However, this strategy may be disputed and can profoundly affect the performance of the learning methods.

This is why another approach is proposed here. In our work, interfaces of known protein-protein complexes are described by collections of small subgraphs taken in a dictionary of elementary patterns, much as transactions in a database of purchases in a supermarket are made of collections of items in a given set of products. This analogy suggests to use data mining techniques in order to detect characteristic regularities in known protein-protein interfaces.

After briefly describing the representation of the protein-protein interfaces, section 2 describes in a generic way the proposed method to analyze whether the interfaces of protein-protein complexes have special properties, and, if yes, which ones. The results obtained using data extracted from the Dockground database [7] are described in section 3. Finally, section 4 discusses the results obtained in light of the overall protein docking problem and opens directions for extensions of this work.

2 The aCID method for the characterization of protein-protein interfaces

In our work, interfaces of known protein-protein complexes are described by collections of *edges*, *triangles* and *tetrahedra* extracted from a geometric representation of proteins called weighted α-complexes [5,8] and completed by additional information about the nature of the amino acids involved in each of them. To obtain a reasonable number of descriptors as compared to the amount of available data, we grouped the amino acids with respect to their physico-chemical properties [4] : *Hydrophobic*, *Polar*, *positively charged*, *negatively charged* and *Small* (resp. noted H, P, $+$, $-$ and S). This leads to a repertoire of 120 distinct descriptive items.

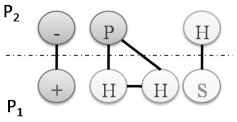


Figure 1. An example of a protein-protein interface in the chosen representation.

In our dataset (see section 3), one typical interface involves between 15 to 50 items, some of them possibly repeated (e.g. figure 1). On average, each interface contains 22 geometrical items, which gives rise to approximately 10,000 items for the whole set of the 459 interfaces studied.

The central question is: *do the interfaces of the known protein-protein complexes present special regularities*?

2.1 Analysis of the frequencies of items

Suppose we observe that a given item, say (SHP), occurs 150 times in all (over the 10,000 items taken altogether) and is present in 50 out of the 459 interfaces. What should we think? Is this feature normal? Mildly surprising? Quite astonishing? One that could be used as a "signature" of a likely interface between proteins? To answer these questions necessitates that expectation under "normal circumstances" be defined (a.k.a. as a "null hypothesis") and that deviations from it can give rise to probability assessments.

In order to compute the probability associated with each item $\mathtt{A}$ (e.g. $(S++)$), one can measure the probability that it would appear as the result of the combination of half-items $\mathtt{A}_i\mathtt{A}_j$ (e.g. $(S++)$ could result from $(S \leftrightsquigarrow ++)$ or from $(+ \leftrightsquigarrow S+)$). In general, given that an event $\mathtt{A}$ can result from pairs of sub-events $\mathtt{A}_i\mathtt{A}_j$, its expected number $n_{\mathtt{A}}$ under the binomial assumption[3] is:

$$\mathbb{E}[n_{\mathtt{A}}] = \sum_{i,j} a_{ij} \cdot p(\mathtt{A}_i) \cdot p(\mathtt{A}_j) \cdot N \tag{1}$$

[3] I.e. independent and identical trials.

where $p(X)$ is the probability of the item X as measured in the interfaces, N is the total number of events and $a_{ij} = 1$ if $A_i = A_j$ or $a_{ij} = 2$ if $A_i <> A_j$. And the variance is given by:

$$Var[n_{\mathtt{A}}] = \big(\sum_{i,j} a_{ij} \cdot p(\mathtt{A}_i) \cdot p(\mathtt{A}_j)\big)\big(1 - \sum_{i,j} a_{ij} \cdot p(\mathtt{A}_i) \cdot p(\mathtt{A}_j)\big)N \quad (2)$$

For instance, suppose again that one is interested in the $(S++)$ item. One would measure the probability of having the semi-item (S), $(+)$, $(++)$ and $(S+)$ which would enable to get: $\mathbb{E}[n_{(S++)}] = 2\big(p(S)\cdot p(++) + p(+)\cdot p(S+)\big)\ N$, where $p(x)$ would be the observed frequency of the semi-item x in all semi-interfaces[4], and N be the number of items in all 459 interfaces, that is 10,000 (the factor 2 comes from $\big(\sum_{i,j} p(\mathtt{A}_i)\cdot p(\mathtt{A}_j)\big) = \big(\sum_{i,i} p(\mathtt{A}_i)\cdot p(\mathtt{A}_i)\big) + 2\big(\sum_{i,j,i<j} p(\mathtt{A}_i)\cdot p(\mathtt{A}_j)\big)$ which reflects the fact that the same item can be obtained with an $\mathtt{A}_i$ coming from either semi-interface).

Note that no combination of semi-items should be considered that lead to items with more than 4 elements (tetrahedra). This must be taken care of in the computation of formula 1 and 2.

2.2 Analysis of the frequencies of the combinations of items

We look for combinations that would be very differently represented than what should be expected under a null hypothesis where the items would be independent. In general, the expected number of a m-combination of m items $\underbrace{\mathtt{A}_i, \mathtt{A}_j, \ldots, \mathtt{A}_k}_{m}$ is:

$$\mathbb{E}[n_{\mathtt{AB}}] = \Pi_{l=i,j,\ldots,k}\ \ a_{i,\ldots,k} \cdot p(\mathtt{A}_l) \cdot N$$

and the variance:

$$Var[n_{\mathtt{A}}] = \big(\Pi_{l=i,j,\ldots,k}\ \ a_{i,\ldots,k} \cdot p(\mathtt{A}_l)\big)\big(1 - \Pi_{l=i,j,\ldots,k}\ \ a_{i,\ldots,k} \cdot p(\mathtt{A}_l)\big) \cdot N$$

with N the number of observed items of size k and $a_{i,\ldots,k}$ the number of permutations of $A_i, \ldots, A_k$.

2.3 Combing the items and combinations

Underrepresented items are not to be retained if the goal is to discover elements that are responsible for the binding of protein-protein complexes. We keep therefore the items (or the combinations of items) of which the observed number in the known interfaces exceeds its expected number by more than twice the standard deviation: $n_X^{obs} \geq \mathbb{E}[n_X] + 2\sqrt{Var[n_X]}$. Under the normal distribution assumption, the probability of observing n_X^{obs} events or more is then less that 2.5% [5]. The choice of this threshold controls the rate of false positive elements (Type 1 error)[6].

In the same spirit, we would rather identify elements that seem well correlated to as large as possible a fraction of all known interfaces. This means both that they are significantly over represented (as detected by the above statistical criterion) and that they intervene in a sufficiently large number of interfaces. The number of interfaces in which a given element X takes part is called the *coverage* of the element and is noted $cov(X)$. A single threshold on the minimal coverage of elements of interest will select the elements that play a role in at least that many interfaces or fraction of the interfaces. For instance, in our study, we chose a 5% threshold for the minimal coverage of elements to be considered for further analysis.

[4] The term *semi-interface* (possibly associated with a subscript) denotes the half belonging to one protein in an interface.

[5] Strictly speaking, the probability of measuring n_X^{obs} outside the range $[\mathbb{E}[n_X] \pm 1.96\sqrt{Var[n_X]}]$ is less than 5%. For symmetry reasons, $p\big(n_X^{obs} \geq \mathbb{E}[n_X] + 1.96\sqrt{Var[n_X]}\big) < 2.5\%$. This is also known as the p-value.

[6] For instance, if one keeps all items with $n_X^{obs} \geq \mathbb{E}[n_X] + \sqrt{Var[n_X]}$, then there is a 16% chance that this happened under the null hypothesis.

2.4 Measuring the spread of the elements and its atypical character

It is also informative to know if a given element (an item or a combination of items) tends to occur in a widespread fashion among the interfaces or, on the contrary, in a concentrated way. In the former case, this might indicate a necessary ingredient in at least one type of bonds between proteins. In the latter case, this could be interpreted as the sign of a kind of autocatalytic reaction that favors the co-occurrence of a same element inside interfaces. Either way, one must be able to measure to which degree an element is more widespread or more concentrated than normal. We therefore propose to *compare the measured coverage of elements with their expected coverage*.

The coverage of an element is easily computed from the database of known instances. The computation of its expected coverage, on the other hand, requires some caution. Suppose that a given element has been observed to occur n times. The idea is to calculate the number of different interfaces among I (e.g. 459) that can receive at least one element when n elements of the same type are drawn independently within N elements.

Suppose that the average number of elements in each interface is $K = N/I$, and let k be the number of a given element in a given interface. Then k is the size of the intersection between a set of n elements drawn independently from N and a set of K elements also drawn independently from N. The hypergeometrical equation gives:

$$p(k) = \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}}$$

In particular, the probability of having a void interface (w.r.t. the element of interest) is:

$$p(0) = \frac{\binom{n}{0}\binom{N-n}{K-0}}{\binom{N}{K}} = \frac{\binom{N-n}{K}}{\binom{N}{K}}$$

And the expected number of *non void* interfaces is:

$$I \cdot \overline{p(0)} = I \cdot \left(1 - \frac{\binom{N-n}{K}}{\binom{N}{K}}\right)$$

3 Results on the protein-protein interfaces

Item selection

459 protein-protein complexes were taken from the PDB database [7,2], the existing total number at the time of the first experiments, and have been described as explained above using 120 different items. For each of these items, the total number of occurrences and the coverage have been measured, while the expected number of occurrences (with standard deviation) and the expected coverage have been computed. The `aCID` method then automatically extracted the items satisfying the three criteria:

- *Overrepresentation.* $n_X^{obs} \geq \mathbb{E}[n_X] + C\sqrt{Var[n_X]}$, with $C = 2$ when selecting items, and $C = 1$ or $C = 2$ when selecting patterns.
- *Minimum coverage*. A threshold of 5% or 7% was used for the selection of patterns. None was used when selecting items.
- *Difference with expected coverage*: $cov(X) > \mathbb{E}[cov(X)]$.

+-	SSP	SHP	SPP	SP-	S+-	HHH	HHP	HH+	HP+	H+-	PP+	P+-	++-

Table 1. Selected items

It is noteworthy (see table 1) that items known as poor candidates such as: $--$, $++$, $---$, $+++$ have been rejected. On the other hand, items corresponding to mildly hydrophobic or strongly hydrophobic elements have been retained, such as HHH, HHP, $HH+$, as well as electrically charged elements such as $+-$, $S+-$, $H+-$, $P+-$, $++-$. All of these items are indeed expected to play a role in protein-protein interfaces since they tend to favor stable conformations.

Pattern selection

The same analysis was carried over for the combinations of items, including *doublets*, *triplets*, and *quadruplets* (no *quintuplet* were found to satisfy the selection criteria). In order to test the robustness of the results, selection was carried out using $C = 2$ and $C = 1$ for the overrepresentation criterion and a minimal coverage threshold of 5%.

Doublets	$\boldsymbol{SSP/SSP}$, $\boldsymbol{SSP/SPP}$, $\boldsymbol{SP-/SP-}$, $\boldsymbol{S+-/S+-}$, $\boldsymbol{S+-/++-}$, $\boldsymbol{HHH/HHH}$, $\boldsymbol{HHH/HHP}$, $\boldsymbol{HHH/HH+}$, $\boldsymbol{HHH/H+-}$, $\boldsymbol{HHP/HHP}$, $\boldsymbol{HHP/HP+}$, $\boldsymbol{HH+/HH+}$, $\boldsymbol{HP+/HP+}$, $\boldsymbol{H+-/H+-}$, $\boldsymbol{H+-/++-}$, $+-/+-$, $+-/++-$, SHP/SHP, $S+-/P+-$, $HHP/HH+$, $HH+/HP+$, $HH+/P+-$, $HH+/++-$
Triplets	$\{\boldsymbol{S+-/S+-/S+-}$, $\boldsymbol{++-/H+-/S+-}$, $\boldsymbol{HH+/HHH/HHH}$, $\boldsymbol{HH+/HH+/HHH}$, $\boldsymbol{H+-/H+-/H+-}$ $\{+-/+-/+-$, $+-/HH+/HH+$, $+-/H+-/H+-$, $SHP/SPP/SSP$, $SHP/SHP/SHP$, $H+-/H+-/S+-$, $HH+/HHH/HHP$, $H+-/HHH/HHP$, $H+-/HH+/HHH$, $H+-/H+-/HH+$, $H+-/H+-/HP+$, $H+-/HH+/HH+\}$
Quadruplets	$\{\boldsymbol{H+-/HH+/HHH/HHH}$, $+-/HH+/HHH/HHH$, $SHP/SPP/SSP/SSP$, $SHP/SHP/SHP/SHP$, $++-/H+-/S+-/S+-$, $H+-/HH+/HH+/HHH$, $HHP/SHP/SHP/SHP$, $HH+/HHH/HHP/HHP$, $H+-/HHH/HHP/HHP\}$

Table 2. Items that are over-represented (C=2 (bold), and C=1) and cover at least 5% of the interfaces. Results for C=1 are a superset of the results for C=2.

Regarding the *doublets* and the *triplets*, one can notice a slight overrepresentation of the groups with amino acids belonging to the hydrophobic group H. In general, however, the items are paired according to global properties. Two groups of patterns emerge. One with a high proportion of hydrophobic amino acids H, the other with opposite charges + and -. Only in one instance these properties are found together: $HH+/++-$.

As for the *quadruplets*, it is noticeable that hydrophobic amino acids are predominant. The electric charges + and - equilibrate each other, and there is a positive charge + left. The groups with hydrophobic amino acids takes over.

4 Discussion and future work

Protein docking introduces very challenging problems. In this work, we used a low-resolution geometrical description of protein-protein interfaces and a data mining approach combined with a null hypothesis criterion. This discovery method can be applied in every contexts where the representation of the instances involves counts of patterns taken in a dictionary that is not too large wrt. the number of instances. Furthermore, it naturally adapts to the discovery of disjunctive concepts i.e. different causal processes. Finally, there is no need for constructing artificial decoys.

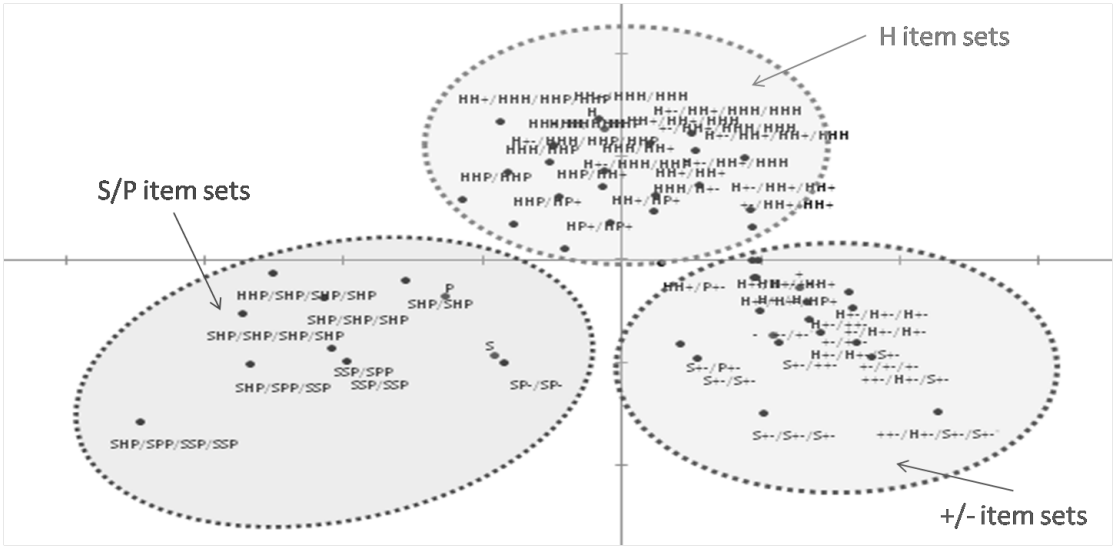


Figure 2. Results of the PCA analysis of the extracted item sets.

Applied to the data set of 459 protein-protein complexes taken from the Dockground database, the `aCID` method selected items and the combinations thereof that point out to the importance of the hydrophobic amino acids and the association of amino acids of opposite charges. The findings are aligned with what is known about protein-protein complexes. Moreover, the results are robust against variations in the grouping of the amino acids into five groups (S, P, H, + and -) and changes in the threshold for selecting significant patterns. The value of the alpha parameter for the alpha-shapes should, however, play a much more important role. This remains to be systematically studied.

References

[1] R.P. Bahadur, P. Chakrabarti, F. Rodier and J. Janin, A Dissection of Specific and Non-specific Protein-Protein Interfaces, *J. Mol. Bio.*, 336, 2004.

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

[3] J. Bernauer, J. Azé, J. Janin and A. Poupon, A new protein-protein docking scoring function based on interface residue properties, *Bioinformatics*, 23:555-562, 2005.

[4] M. Betts and R. Russell, Amino acid properties and consequences of substitutions, *Bioinformatics for Geneticists*, M.R. Barnes, John Wiley and Sons, pp. 291-315, 2003.

[5] F. Cazals, J. Giesen, M. Pauly and A. Zomorodian, Conformal Alpha Shapes, *In proceedinfs of Eurographics Symposium on Point-Based Graphics*, 2005.

[6] P. Chakrabarti and J. Janin, Dissecting protein-protein recognition sites, *Proteins: Structure, Function, and Bioinformatics*, 47:334-343, 2002.

[7] D. Douguet, H.-C. Chen, A. Tovchigrechko and I. A. Vakser, Dockground resource for studying protein-protein interfaces, *Bioinformatics*, Oxford University Press, Oxford, UK, 22:2612-2618, 2006.

[8] H. Edelsbrunner, Weighted alpha shapes, Technical report, University of Illinois at Urbana-Champaign, Champaign, IL, USA, 1992.

[9] D. Reichmann, O. Rahat, S. Albeck, R. Meged, O. Dym and G. Schreiber, From The Cover: The modular architecture of protein-protein binding interfaces, *PNAS*, 102:57-62, 2005.

[10] G. R. Smith and M. J. Sternberg, Prediction of protein-protein interactions by docking methods, *Curr Opin Struct Biol.*, Feb, vol. 12, num. 1, pp. 28-35, 2002.

FUNGIpath: a new tool for analysing the evolution of fungal metabolic pathways

Sandrine Grossetête, Bernard Labedan, Olivier Lespinet

Institut de Génétique et de Microbiologie, UMR 8621,
Université Paris Sud, Bâtiment 400, 91405 Orsay Cedex

{sandrine.grossetete,bernard.labedan,olivier.lespinet}@igmors.u-psud.fr

Abstract: *FUNGIpath is a new tool dedicated to perform in-depth analysis of fungal metabolic pathways. It is freely accessible at http://www.fungipath.u-psud.fr. FUNGIpath consists in a collection of orthologous groups of proteins that have been predicted using complementary methods of detection and further mapped on KEGG and MetaCyc pathways. It allows an easy comparison of the primary and secondary metabolisms afforded by the different fungal species present in the database with the possibility to assess the level of specificity of various pathways at different taxonomic distances. As more and more fungal genomes are expected to be decrypted in the next years, this tool is expected to help to progressively reconstruct what were the primary and secondary metabolisms of the ancestors of the main branches of the fungi tree and to understand how these ancestral fungal metabolisms evolved to various specific derived metabolisms.*

Keywords: Metabolism, evolution, fungi.

1 Introduction

Fungi constitute one of the eukaryotic taxonomic group that present today (April 2009) the highest number of species for which the complete sequence of the nuclear genome has been published and is available to the scientific community (26 genomes according to [1]). This relative abundance is mainly due to their moderate genome size, and to the fact that several species have been model organisms for fundamental, medical, or agronomical and industrial studies (e.g. *Saccharomyces cerevisiae, Candida albicans, Yarrowia lipolytica*).

Therefore, fungal genomes appear today to be a suitable material for large-scale comparative studies. Indeed, several teams have already performed extensive comparison of a few fungal genomes to predict clusters of orthologous groups of proteins, which can be accessed by tools such as OrthoDB [2] or e-Fungi [3]. Such approaches open the way to study the evolution of fungal genomes [4].

However, information about the metabolism of fungi is presently rather scarce and heterogeneous in major public databases. Although we found a moderate or low amount of data in dedicated databases such as KEGG [6] or MetaCyc [7], there are almost no data on fungi metabolism in Swiss-Prot [5] with the noticeable exception of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (data not shown). In addition, beside a preliminary attempt to identify enzymes in pathogenic fungi for a limited number of metabolic pathways [8], there is presently no tool allowing performing large-scale analysis of fungal metabolism.

Here, we describe FUNGIpath that is, to our knowledge, the first tool allowing to mining genomic data in order to perform in-depth analysis of fungi metabolism. This new tool presents two efficient features: it is based on several complementary approaches combining to define reliable groups of orthologous genes and it allows mapping these groups on every pathway that are available in the KEGG [6] and MetaCyc [7]

databases.

2 Organizing Metabolic Data by Comparing Fungal Genomes

Primary (sequences and pathways) and secondary data (orthologous group) were assembled in a database that is made freely available to the community through FUNGIpath, a user-friendly website implemented in PHP, HTML and Javascript.

2.1 Primary Data

Sequences data are summarized in two tables, one describing genome informations and the other one listing the amino acid sequences encoded by the genomes of 20 fungal species: *Aspergillus nidulans*, *Aspergillus oryzae*, *Batrachochytrium dendrobatidis*, *Chaetomium globosum*, *Coprinus cinereus*, *Fusarium oxysporum*, *Laccaria bicolor*, *Magnaporthe bicolor*, *Mycospherella graminicolla*, *Neurospora crassa*, *Phycomyces blakesleeanus*, *Podospora anserina*, *Puccinia graminis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Sclerotinia sclerotiorum*, *Stagonospora nodorum*, *Trichoderma reesei*, *Ustilago maydis*, *Yarrowia lipolytica.*

The complete list of genome sources and url files are available on supplementary data table 1. For each genome, we removed sequences that are 100% identical. Supplementary data Table 1 provides sources for the respective genomic data used in FUNGIpath.

Metabolism data downloaded from either KEGG (6) or MetaCyc (7) and enriched with predicted annotations are organized in several tables to speed up data access time.

2.2 Predicting Orthologs

Different methods have been published to predict orthologs but none of them appears completely reliable since they poorly overlap (supplementary data, table 2). Thus, we found necessary to use independent methods to collect as many potential orthologs as possible. Moreover, exploring several methods raised the probability to have a consistent group, corresponding to their overlapping. Accordingly, we are using three different and complementary approaches based on similarity searches and another one based on the analysis of phylogenetic trees.

First, we adapted two methods already published with their respective default parameters: OrthoMCL [9] allows defining consistent groups of orthologs that are strongly related. Inparanoid [10] permits to differentiate orthologs and inparalogs (genes recently duplicated after the last speciation event) in pairwise comparison of all genomes. Moreover, the classical Best Reciprocal Hits (BRH) approach has been entirely automated by a Perl script. To improve the definition of orthologs we filtered the BLAST [11] results by specifying two parameters, the alignment percent and the score ratio. Dividing the alignment length of each aligned sequence by its total length permits to avoid local conservation. The score ratio is computed by dividing the crude BLAST score obtained when aligning sequence 1 against sequence 2 by maximum BLAST score, i.e. BLAST score obtained when sequence 1 is aligned against itself. We keep only results with score ratio superior to 0.2 and alignment percent superior to 60%.

These different methods based on sequence similarity allow to get more or less stringent clusters of orthologous genes depending if we used single (e.g. Inparanoid) or multiple (BRH) linkage.

Beside these methods based on similarity approaches, we also used a phylogeny approach to get orthologous groups using the automatic tree analysis previously developed by Lemoine et al. [12]. We first build families of homologous protein detected by BLASTP [11] with the following requirements: an E-value less than 0.001 and an alignment percent larger than 70 % of the length of the shorter sequence of the aligned pair. For each family, a multiple alignment was built with Muscle [13], and the deduced phylogenetic tree was reconstructed with PhyML [14]. The program Retree from Phylip package [15] was further used to root the tree in order to distinguish orthologs and paralogs with the automatic tree analysis [12].

Once we got the orthologous groups, we compared groups obtained by the different methods and merged groups that overlap (supplementary data Table 2). To help the user to evaluate the reliability of the predictions we computed a confidence score based on the number of methods that found independently the same group. We suppose that a group found by several methods is more reliable than a group found only by one of the available methods (see supplementary data for score computation). We obtained orthologous group with group size ranging from 2 to 2694 sequences. As the homogeneity of the largest groups is most propably doubtful, we kept only orthologous groups with a score superior to 1.5. Such a threshold value, helped to limit the group size to a maximum of 400 sequences.

2.3 Reconstructing Pathways

Annotating the putative enzymatic activities

Once the orthologous groups have been defined, we attemped to predict functionnal annotation by using an HMM approach. For each orthologous group, we built the corresponding HMM profile with hmmbuild [16] and used hmmsearch [16] to search its similarity with sequences that display an enzymatic annotation in Swiss-Prot [5]. The annotation was transferred to the orthologous group analyzed, if the E-value of the best hit obtained is lower or equal to 10^{-80}.

This annotation assigned 843 different EC numbers to 1261 groups of orthologous proteins. Among these groups, our HMM approach contribute to annotate 360 groups (29%) which the belonging sequences did not get any annotation in the Swiss-Prot database [5]. Nearly one half (396) of the EC numbers is present in all genomes, and 90% (764) of the assigned EC numbers are found in at least 50% of the genomes.

Assembling the putative EC numbers in pathways

Once the different putative orthologs have been annotated as described above, we used them to exhaustively reconstruct the different metabolic pathways in fungi. To do that, we used two reliable public databases, KEGG [6] and MetaCyc [7] that differ in their way to define pathways.

Useful information was extracted from the reaction file generated by KEGG [6] and the corresponding GIF maps were downloaded. BIOPAX files defined in MetaCyc [7] were downloaded and we generated automatically the corresponding map pictures by directed graph building. Accordingly, we collected 151 pathways in KEGG and 1143 pathways in MetaCyc that define mainly anabolic and catabolic ways.

3 Querying FUNGIpath and exploring pathways

FUNGIpath (http://www.fungipath.u-psud.fr) has been designed to allow studying fungi metabolism by performing various queries on our database.

FUNGIpath allows checking and visualizing the conservation of pathways between different fungi. We can make such a search using several ways: one can start from a defined EC number (Fig. 1), from a known pathway (Fig. 2), or by using a user-defined pathway delineated in a simplified BIOPAX format (data not shown).

Searching a specific EC number allows assessing the level of conservation of this EC number in each taxonomic group and to directly access to all the pathways in which this EC number is involved (Fig. 1).

For instance, Fig. 1 shows that EC 3.5.1.4 corresponds to an amidase (Acymlamide amidohydrolase that cleaves carbon–nitrogen bonds in amides) that is very well conserved in all fungi and is involved in at least 6 different pathways in both KEGG and MetaCyc databases. Moreover, we found that this EC number has been assigned to several distinct orthologous groups (data not shown).

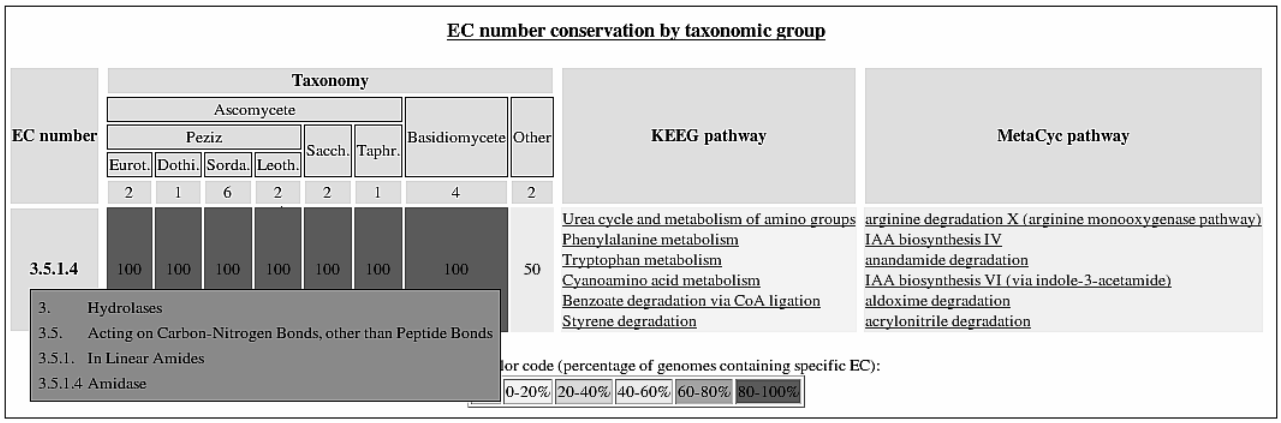


Figure 1. Exploring pathways using a specific EC number (3.5.1.4). The level of occurrence in the different species belonging to the different taxonomic groups of fungi is indicated with a color code (scale from white to red) in the taxonomy column. The respective lists of the pathways that contain the requested EC number are indicated in the KEGG and MetaCyc columns, respectively. Note that the pathway names are different in both databases.

Fig. 2 illustrate the case of the 'Biotin metabolism' (KEGG database), the results displayed when a pathway is queried.

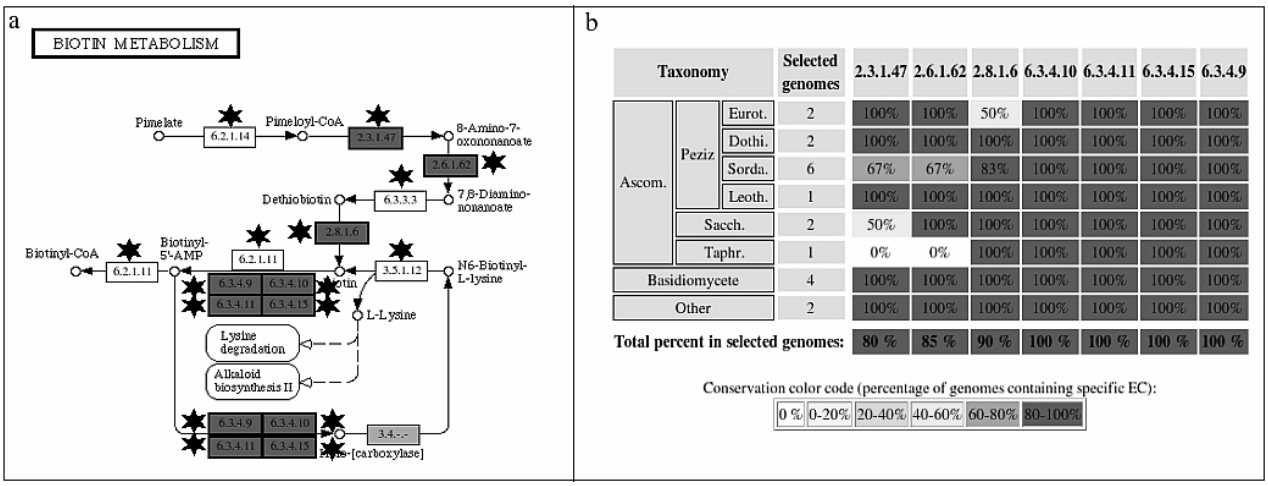


Figure 2. Exploring pathways using a specific pathway name. **a:** The 'Biotin metabolism' pathway is displayed using the KEGG map. For each EC a color code indicate the level of conservation. The black star indicates the EC numbers that are specific to the 'Biotin metabolism' **b:** The table lists the percentage of conservation of this pathway with the same color code (white to red) in each species belonging to the different taxonomic groups of fungi. Moreover, the relative presence in all 20 fungi is given in the last line of this table, this value defining the color used in the KEGG map, in. 2a.

To facilitate the analysis of these results, a color code has been associated to the conservation level of EC numbers. Results are presented both as a KEGG gif map (Fig. 2a) and summarized in a table (Fig. 2b). For each EC number the corresponding orthologous groups of proteins can easily be accessed by using the genomes feature table (not shown) and can be downloaded for further studies.

4 Discussion

4.1 Improving functional annotation of fungi genomes

To challenge the validity of our predictions of functional annotation, we compared as a control all our EC numbers predictions for the yeast *S.cerevisiae* with 4 different curated public databases (KEGG [6], MetaCyc [7], Swiss-Prot [8], and SGD [17]). The results are displayed in Table 1. We call ID-EC the unique

pair formed between an ID and its EC (or one of its EC). Thus, an EC can belong to several pairs, if several IDs have the same annotation. Likewise, an ID can be present in several pairs, if the ID gets several EC numbers in case of multifunctional proteins.

Database	Total number of ID-EC in FUNGIpath	Number (percent) of identical predicted ID-EC	Number of different ID-EC	Digit position which is different			
				1st d.	2nd d.	3rd d.	4th d.
KEGG	1062	843 (79,4%)	47	2	2	7	36
METACYC	523	409 (78,2%)	68	16	3	7	42
SGD	512	408 (79,7%)	41	9	2	3	27
SP	1114	1030 (92,5%)	40	2	1	5	32
Prediction	1299						

Table 1. Comparison of enzymatic data for *S. cerevisiae* between 4 databases and our prediction

According to our predictions made from orthologous groups we get 1299 ID-EC for the yeast. The observed high overlapping with Swiss-Prot figures is not surprising since we used mainly this database for our prediction. On the other hand, our prediction corresponds to nearly 80% of the EC numbers found in the three other databases. In addition, most of the IDs pairing with different EC numbers diverge only at the level of the last digit. Thus, the reliability of the predictions of our group of orthologous proteins and of their enzyme function appears to be comparable with that of the independently curated public databases.

Moreover, we compared the functional annotations for 12 species between KEGG [6] and FUNGIpath (Supplementary data Table 3). More than 75% of KEGG data are found in FUNGIpath. However, contrarily to KEGG, we work only with complete EC numbers. Thus, the comparison of the incomplete KEGG EC numbers with our data allows to recover some of them. For instance, in the case of *Schizosaccharomyces pombe*, the total number of annotated enzymes are very close (1267 EC numbers in KEGG and 1231 ones in FUNGIpath), but among the 258 EC numbers that are incomplete in KEGG, 62 have been completed in FUNGIpath. Thus, FUNGIpath is a efficient tool for functional annotation of fungi and thus for studying their metabolism.

4.2 Studying annotation and evolution of metabolism in fungi

One of the main problems encountered when trying to reconstruct entire pathways from orthology data is the occurrence of missing data. The absence of an EC number (*orphan metabolic activities* [18]), may be due to a too low percent identity of the corresponding amino acid sequence or to its replacement by another protein. Alternatively, the simultaneous absence of several EC numbers that belong to a specific pathway most likely suggests that this entire pathway is not present in the studied species. However, one cannot dismiss the hypotheses that this absence is simply due to a major annotation problem or to the replacement of this pathway by an alternate one.

For instance, Fig. 2 shows that most of the EC numbers involved in the 'Biotin metabolism' defined by KEGG [6], are found in our database and appear specific to this pathway. However, this Biotin metabolism pathway appears to be incomplete in many fungi since several of its specific enzymatic activities are not found such as 3.5.1.12, 6.3.3.3 and 6.2.1.11 (white EC with black stars in Fig. 2a). We can suppose that either these EC numbers exist in fungi but they are presently not detectable, or fungi use other EC numbers to catalyse these reactions. Of the 11 EC numbers involved in the 'Biotin metabolism', four (6.3.4.9, 6.3.4.10, 6.3.4.11, 6.3.4.15) are found in all the selected species and one (2.8.1.6) in all species excepted one genome (*A.orizae*). Two EC numbers (2.3.1.47, 2.6.1.62) are absent in several genomes (*M. grisea, N. crassa, S. pombe* plus *S. cerevisiae* for 2.3.1.47). These genomes may use an alternative way to realize the first steps of the 'Biotin metabolism'.

5. Conclusion

FUNGIpath appears a reliable tool that helps to analyse the metabolism of fungi. It will be especially

useful to annotate newly-sequenced genomes.

Moreover, FUNGIpath allows an easy comparison of the respective metabolisms afforded by the different taxons. For instance, 163 EC numbers are found uniquely in Ascomycetes (data not shown) and may help to delineate the metabolic specificities of the common ancestor to this group.

As more and more genomes are expected to be decrypted in the next years, FUNGIpath we will be especially usefull to progressively reconstruct what were the primary and secondary metabolisms of the ancestors of the main branches of the fungi tree and to understand how these ancestral fungal metabolisms evolved to various specific derived metabolisms.

Acknowledgements

We are grateful to Philippe Silar for his help during the designing process of the web site and his helpful comments about this work. The computations were performed on the MIGALE platform (INRA, Jouy-en-Josas, France). SG is a PhD student supported by a doctorant CNRS fellowship.

References

[1] K. Liolios, K. Mavrommatis, N. Tavernarakis, N.C. Kyrpides, The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res., 36, D475-479, 2007.

[2] E.V. Kriventseva, N. Rahman, O. Espinosa, E.M. Zdobnov, OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res., 36, D271-275, 2008.

[3] C. Hedeler, H.M. Wong, M.J. Cornell, I. Alam, D.M. Soanes, M. Rattray, S.J. Hubbard, N.J. Talbot, S.G. Oliver, N.W. Paton, e-Fungi: a data resource for comparative analysis of fungal genomes. BMC Genomics, 8:426, 2007.

[4] D.M. Soanes, I. Alam, M. Cornell, H.M.Wong, C. Hedeler, N.W. Paton, M. Rattray, S.J. Hubbard, S.G. Oliver, N.J. Talbot, Comparative Genome Analysis of Filamentous Fungi Reveals Gene Family Expansions Associated with Fungal Pathogenesis. PLoS ONE, 3, 6:e2300, 2008.

[5] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivany, R.D. Appel, A. Bairoch, Nucleic Acids Res., 31:3784-3788, 2003.

[6] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res., 27:29-34, 1999.

[7] R. Caspi, H. Foerster, C.A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S.Y. Rhee, A.G. Shearer, C. Tissier, T.C. Walk, P. Zhang, P.D. Karp, The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res., 36:623-31, 2008.

[8] P.F. Giles, D.M. Soanes, N.J. Talbot, A relational database for the discovery of genes encoding amino acid biosynthetic enzymes in pathogenic fungi. Comp Funct Genomics. 4(1):4-15, 2003.

[9] L. Li, CJ. Jr. Stoeckert, D.S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res., 13, 2178-89, 2003.

[10] M. Remm, C.E. Storm, E.L. Sonnhammer, Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol., 314, 1041-52, 2001.

[11] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. J Mol Biol., 215:403-10, 1990.

[12] F. Lemoine, O. Lespinet, B. Labedan, Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. BMC Evol Biol., 7, 237, 2007.

[13] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res., 32:1792-7, 2004.

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

[15] J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2005.

[16] SR. Eddy, Hidden Markov models. Curr Opin Struct Biol. 6(3):361-5, 1996.

[17] JM. Cherry, C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, RK. Mortimer, D. Botstein, Genetic and physical maps of Saccharomyces cerevisiae. Nature 387(6632 Suppl):67-73,1997.

[18] L. Chen, D. Vitkup, Predicting genes for orphan metabolic activities using phylogenetic profiles, Genome Biol., 7:R17, 2006.

Detecting Network Motifs by Local Concentration

Etienne Birmelé[1]

Laboratoire Statistique et Génome, UMR CNRS 8071, INRA 1152
Tour Evry 2, 523 place des Terrasses de l'Agora 91000 EVRY France
etienne.birmele@genopole.cnrs.fr

Abstract: *Biological networks exhibit small over-represented subgraphs, called motifs, some of which are known to have a biological function. Several algorithms exist to detect motifs, most of them being based on time-consuming simulations or leading to many false positives. We propose an efficient and conservative procedure to detect network motifs and apply it on the Yeast gene regulation network.*

Keywords: Networks, Motifs, Concentration inequalities.

1 Introduction

Recent work indicates that biological networks show recurrent small patterns, called *network motifs* [1]. They can be thought of as small units of given function from which the networks are built: it is then quite natural to ask which are the small patterns that are over-represented in given networks. Many attempts were made to answer that question. Some of them [1] compute a huge number of random networks with the same degree distribution as the biological one, but are not tractable for motifs on more than four vertices. Others [2,3] rely on a sampling algorithm and the calculation of a Z-score. Nevertheless, the distribution of the number of small subgraphs is more heavy-tailed than a gaussian and therefore those methods have no control on false positives.

To avoid simulations, one has to define a probabilistic model of random graph generation and to compute the p-value of the observed number of subgraphs in that model. Mixture models [4,5] were shown to be relevant as they give rise to graphs depending on independent Bernoulli trials but which exhibit the heterogeneity observed in biological networks. Picard and al. [6] look for motifs by taking the latter model as the null model. As they can't compute exactly the law of the number of small subgraphs, they propose to fit the best possible Polya-Aeppli distribution to the subgraph distribution and to take the p-value of that approximate distribution.

As pointed out in [1], another issue is that a motif can appear as over-represented because it contains an over-represented sub-motif, which is in fact the biological relevant structure. Moreover, Dobrin and al. [7] show that the motifs in the yeast transcriptional regulatory network aggregate. For both reasons, we take another approach: we consider a small subgraph $\mathbf{m}$ and a subgraph $\mathbf{m}'$ of $\mathbf{m}$ obtained by deleting one vertex. We then define $\mathbf{m}$ to be a motif with respect to $\mathbf{m}'$ if there exist an occurrence of $\mathbf{m}'$ in the network such that the number of occurrences of $\mathbf{m}$ sharing the given occurrence of $\mathbf{m}'$ is overrepresented.

Furthermore, to avoid a bias due to the use of an approximate distribution, we will determine an upper bound of the real p-value. To do so, we will show that the number of motifs aggregating on a given submotif is highly concentrated around its expectation. Our method is thus conservative, but ensures a low rate of false positive. Moreover, we apply our method to the Yeast gene interaction network, showing that all the known motifs are found again and with supplementary informations.

2 Notations and definitions

In the following, we will consider a directed network G of vertex set V and edge set E. We suppose that there are no multiple edges in the same direction but opposite edges between two vertices and self-loops are allowed. For any vertex set $U \subset V$, we denote by $G[U]$ the *induced* subgraph of G on the set U, that is the graph of vertex set U where each edge is present if and only if it is present in G.

Let $\mathbf{m}$ be a small graph on k vertices, which we want to determine if it is overrepresented. Let s be one of the vertices of $\mathbf{m}$ and denote by $\mathbf{m}'$ the graph on $k-1$ vertices obtained by deleting s in $\mathbf{m}$. Figure 1 shows an example of $\mathbf{m}$ and $\mathbf{m}'$.

The counting random variables of interest are the following: $N(\mathbf{m})$ is the number of occurrences of the motif in G, and, for every set U of $k-1$ vertices corresponding to an occurrence of $\mathbf{m}'$, $N_U(\mathbf{m})$ denotes the number of occurrences of $\mathbf{m}$ which are extensions from the occurrence of $\mathbf{m}'$ (cf Figure 1). If U does not correspond to an occurrence of $\mathbf{m}'$, we set $N_U(\mathbf{m}) = 0$.

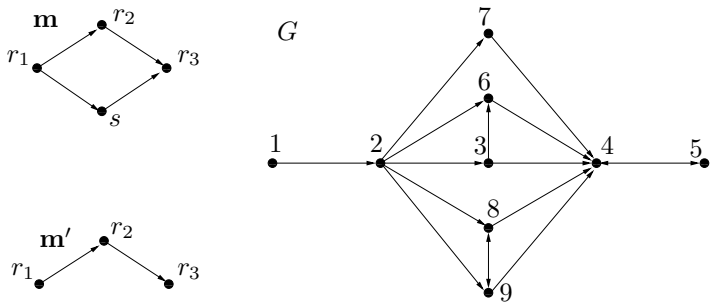


Figure 1. For $U = \{2, 3, 4\}$, an occurrence of $\mathbf{m}'$ is present on U and $N_U(\mathbf{m}) = 3$. Indeed, one obtains valid extensions of $\mathbf{m}'$ to $\mathbf{m}$ by adding to U the vertices 7, 8 or 9. Adding the vertex 6 does not give rise to a valid extension because of the edge between 3 and 6.

The random graph model we consider is the mixture model with fixed classes. In that model, the n vertices are spread into Q known classes.

We consider a matrix $\Pi = (\pi_{qr})_{1 \leq q,r \leq Q}$ which gives the connection probabilities between classes and, for each pair (i, j) of vertices, the edge $\vec{ij}$ is present with probability π_{qr}, q and r being the respective classes of i and j.

To estimate the partition of the graph, including the number of classes to choose, and the corresponding matrix Π, we use the algorithm developped by Latouche and al. [5], assigning each vertex to its most probable class.

3 A Concentration Inequality to detect Network Motifs

To determine if $\mathbf{m}$ is a motif, our strategy is to look for an occurrence of $\mathbf{m}'$ on a set U such that $N_U(\mathbf{m})$ is much larger than expected. To do this, we show that $N_U(\mathbf{m})$ is highly concentrated around its mean using the following concentration inequality [8]:

THEOREM 3.1. *Let the random variables $X_1, \ldots, X_n$ be independent, with $0 \leq X_k \leq 1$ for each k, and let $S_n = \sum_{k=1}^{n} X_k$. Then, for every $t > 0$,*

$$\mathbb{P}\big(\frac{S_n - \mathbb{E}S_n}{\mathbb{E}S_n} > t\big) \leq e^{-((1+t)\ln(1+t)-t)\mathbb{E}S_n}$$

We will here just enounce our results and give a sketch of the proofs, without detailed notations and precise mathematical justifications. The interested reader will find them in [9].

3.1 First step: Local Overrepresentation

Let us consider a set U of $k-1$ vertices corresponding to an occurrence of $\mathbf{m}'$. For each vertex $v \notin U$, we define ext_U^v as the indicator of the fact that adding v to U leads to an occurrence of $\mathbf{m}$. For instance, in Figure 1, $ext_U^9 = 1$ but $ext_U^1 = 0$. Let Ext_U be the mean of the valid extensions, that is $Ext_U = \mathbb{E}(\sum_{v\notin U} ext_U^v)$.

It is then straightforward to see that:

- $N_U(\mathbf{m}) = \mathbb{I}_{G[U]\sim\mathbf{m}'} \sum_{v\notin U} ext_U^v$ where $G[U] \sim \mathbf{m}'$ denotes that U is an occurrence of $\mathbf{m}'$.
- Each ext_U^v depends only on the edges between U and v and thus the variables $\big(ext_U^v\big)_{v\notin U}$ are independent.

Therefore, we can apply Theorem 3.1 to obtain, for every $t > 0$,

$$\mathbb{P}\big(\frac{N_U(\mathbf{m}) - Ext_U}{Ext_U} > t \quad | \quad G[U] \sim \mathbf{m}'\big) \leq e^{-((1+t)\ln(1+t)-t)Ext_U}$$

Taking also the case when U does not correspond to an occurrence of $\mathbf{m}'$ into account yields

$$\mathbb{P}\big(\frac{N_U(\mathbf{m}) - Ext_U}{Ext_U} > t\big) \leq \mathbb{P}(G[U] \sim \mathbf{m}')e^{-((1+t)\ln(1+t)-t)Ext_U} \tag{1}$$

We thus obtain an exponentially decreasing local bound for the p-value of many occurrences of $\mathbf{m}$ sharing the same subgraph $\mathbf{m}'$ located on U.

3.2 Second step: A Global Statistic to detect Motifs

Equation 1 gives a local p-value, but as the number of possible positions U is growing as n^{k-1}, there is a problem of multiple testing. To overcome it, we have to build a statistic characterizing any local overrepresentation somewhere in the graph.

Let h be the function defined on $[0, +\infty[\times]0, +\infty[$ by

$$h(X,Y) = \begin{cases} 0 & \text{if } X \leq Y \\ X\ln(\frac{X}{eY}) + Y & \text{else} \end{cases}$$

Note that for a fixed value of Y, $h_Y : X \to h(X,Y)$ is an increasing function, growing asymptotically as $X\ln(X)$, as illustrated in Figure 2.

We can show that Equation (1) can be rewritten as follows:

$$\forall t > 0, \mathbb{P}\big(h(N_U(\mathbf{m}), Ext_U) > t\big) \leq \mathbb{P}(G(U] \sim \mathbf{m}')e^{-t}$$

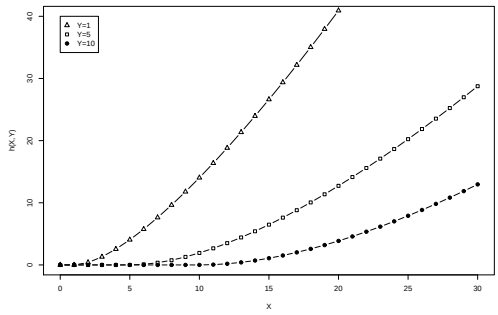


Figure 2. Curves of the function h for fixed values of Y

Noting that $\{\max_U(h(N_U(\mathbf{m}), Ext_U)) > t\} = \bigcup_U \{h(N_U(\mathbf{m}), Ext_U) > t\}$ and that the exponential depends no more on U, one can obtain our main result:

THEOREM 3.2.

Let $aut(\mathbf{m}')$ be the number of automorphisms of $\mathbf{m}'$. For every $t > 0$,

$$\mathbb{P}\big(\max_U(h(N_U(\mathbf{m}), Ext_U)) > t\big) \leq aut(\mathbf{m}')\mathbb{E}N_U(\mathbf{m}')e^{-t}$$

We thus obtain a general p-value for the network characterizing a local overrepresentation of $\mathbf{m}$ with respect to $\mathbf{m}'$ somewhere in the network.

The function h is introduced in our statistic in order to give an upper-bound of the real p-value which is as tight as possible. Nevertheless, it can be shown that Theorem 3.2 induces the following result, which is weaker but more explicit:

THEOREM 3.3. *For every $t > 0$,*

$$\mathbb{P}\big(\exists U / N_U(\mathbf{m}) \geq e^2 Ext_U + t\big) \leq aut(\mathbf{m}')\mathbb{E}N(\mathbf{m}')e^{-t}$$

4 Application to the gene regulation network of Yeast

The method described in Section 3 was applied to the gene regulation network of Yeast [10]. That network has 688 vertices and 1078 edges and a directed edge between genes g_1 and g_2 denotes a regulation from gene g_1 on the expression of gene g_2. We don't take the type of regulation into account here, that is if it is an activation or an inhibition. The estimation of the parameters of the model spreads the genes in 5 classes: two of them are small (4 and 8 genes) and correspond to the genes of high degree, two are intermediate (40 and 115 genes) and the biggest one (521 genes) contains most of the vertices of small degree.

Tables 1 shows the unique motif of size three detected with a threshold of 10^{-4}. The column *p-value* contains the upper bound of the real p-value given by Theorem 3.2 for $t = \max_U(h(N_U(\mathbf{m}), Ext_U))$, whereas the *Agglomeration* coefficient is the maximal observed number of extensions of an occurrence of $\mathbf{m}'$.

That motif, called the feed-forward loop, is known (see [11] for a deeper biological insight) to play a role in regulation processes by inducing a delay in the regulation of Z by X: X has first to

activate Y and then X and Y together regulate Z. Our analysis finds this motif again and shows that there are genes X using the same Y to regulate a high number of genes Z (up to 15). On the contrary, there is no gene X using a high number of intermediates Y to regulate a given gene Z.

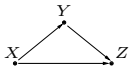
Motif	Deletion class	p-value bound	Agglomeration
	X	$5.11e-3$	3
	Y	3.4	2
	Z	$1.06e-11$	15

Table 1. p-values for the three possible sub-motifs of the feed-forward loop

Table 2 show all the motifs of size four which are found overrepresented with respect to at least one of their submotifs, with a threshold of 10^{-4}.

The motif of size four with the lowest p-value is the bi-fan, that is the first motif shown in table 2. That motif consists on two regulators having impact on two common genes. It was first shown to be over-represented in that network by Milo and al [1] and appears first in all motif detection algorithms.

Nevertheless, our approach gives a supplementary information, that is that it is highly overrepresented with respect to the sub-motif obtained by suppressing one of the regulated genes. In other words, there exist in Yeast co-regulators which co-regulate a high number of genes simultaneously. The agglomeration coefficient shows that they may act on up to 37 common genes. On the other hand, the bi-fan is *not* over-represented with respect to the sub-motif obtained by suppressing one of the regulators, that is there exist no couple of genes that are influenced by a high number of common regulators. In fact, running the algorithm in that case leads to a p-value of $.17$ and an agglomeration coefficient of 4. That asymetry between couples of regulators and couples of regulated genes is well known by biologists but is found again here only by statistical means.

Note also that among the six detected motifs, the third and the two last ones are in fact by-products of the over-representation of the feed-forward loop: they are not overrepresented with respect to their feed-forward loop submotif, that is the one obtained by deleting T.

From a computational point of view, we obtain a significative improvement in terms of running time. Indeed, methods available are of two kinds: either they use a reasonable number of simulations and use a Z-score and are quite rapid but lead to false positives. Or they make use of a huge number of simulated graphs to obtain an empirical p-value but are very time-consuming. The part of our method which takes the most time is in fact the preprocessing by the algorithm of Latouche et al. [5] to estimate the parameters. The concentration part is very rapid as it only needs expectations, which are easy to calculate in mixture models. Comparing our algorithm with the empirical p-value option of the MFinder tool [1] for the graphs of size four in the Yeast network divides the running time by a factor 100 (less than 10 minutes compared to more than 15 hours).

5 Conclusion and Perspectives

We propose to detect network motifs by looking for a local over-representation rather than by studying the total number of occurrences. That approach allows to take into account the subgraphs of the small graphs of interest and thus to study separately the influence of each of the vertices of the found motifs. Comparing with existing methods on the Yeast regulation network, we find again the known motifs, with a deeper biological interpretation and an improvement of the running time.

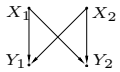
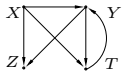
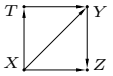
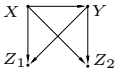
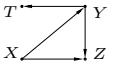
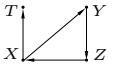
Motif	Deletion class	p-value bound	Agglomeration
	$\{Y_1, Y_2\}$	$1.57e-25$	37
	Z	$3.06e-16$	15
	Z	$2.10e-13$	15
	$\{Z_1, Z_2\}$	$4.52e-12$	14
	Z	$7.70e-12$	15
	Z	$2.00e-10$	15

Table 2. Over-represented motifs of size four

This work will be completed by a simulation study of the loss of precision induced by the upper bound of the p-value and biological applications in order to detect unknown relevant structures. Moreover, the procedure gives in fact local scores to network motifs rather than just a p-value. The biological relevance of those scores is a point to be explored.

References

[1] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, Network Motifs: Simple Building Blocks of Complex Networks, *Science*, 298:824-827, 2002.

[2] N. Kashtan, S. Itzkovitz, R. Milo and U. Alon, Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs, *Bioinformatics*, 20:1746-1758, 2004.

[3] S. Wernicke and F.Rasche, FANMOD: a tool for fast network motif detection, *Bioinformatics*, 22:1152-1153, 2006.

[4] J.-J. Daudin, F. Picard and S. Robin, Mixture model for random graphs, *Statistics and Computing*, 18(2):173-183, 2008.

[5] P. Latouche, E. Birmelé and C.Ambroise, Bayesian methods for graph clustering, *preprint SSB*, 2008.

[6] F. Picard, J.-J. Daudin, M. Koskas, S. Schbath and S. Robin, Assessing the exceptionality of network motifs, *Journal of computational biology*, 15(1):1-20, 2008.

[7] R. Dobrin, Q.K. Beg, A.-L. Barabasi and N. Oltvai, Aggregation of topological motifs in *Escherischia Coli* transcriptional regulatory network, *BMC Bioinformatics*, 5:10, 2004.

[8] C. McDiarmid, Concentration, in M. Habib, C. McDiarmid, J. Ramirez-Alfonsin and B. Reed (eds.), *Probabilistic Methods for Algorithmic Discrete Mathematics*, Springer, pp 195-248, 1998.

[9] E. Birmelé, Detecting network motifs by local concentration, Preprint SSB, http://ssbgroup.fr/preprint.html

[10] Data available on http://www.weizmann.ac.il/mcb/UriAlon/

[11] U. Alon, Network motifs: theory and experimental approaches, *Nature Reviews Genetics*, 8:450-461, 2007.

Meristematic waves, a new approach to root architecture dynamics

Lionel Dupuy[1], Matthieu Vignes[2], Blair McKenzie[1] and Philip J. White[1]

[1] Scottish Crop Research Institute
Invergowrie, Dundee, DD2 5DA, Scotland, UK
lionel.dupuy@scri.ac.uk
[2] SaAB Team, BIA Unit, INRA
Chemin de Borde Rouge, Auzeville, BP 52627, 31326 Castanet-Tolosan Cedex, France
matthieu.vignes@toulouse.inra.fr

Abstract: *During their development, plants must develop efficient root architectures to secure access to nutrients and water in soil. A series of expansion and branching mechanisms fulfils this aim in the proximity of root apical meristems where the plant senses the environment and explores immediate regions of soil. We have developed a new approach to study the dynamics of root meristems in soil, using the relationship between the increase in root length density and the root meristem density. Initiated at the seed, the location of root meristems was shown to propagate, wave-like, through the soil, leaving behind a permanent network of roots for the plant to acquire water and nutrients. Models highlighted that the morphologies of the waves of meristems are inherent to individual root developmental processes, namely expansion, lateral root initiation and gravitropic responses. The meristematic wave observed on data collected on barley might be a more general and fundamental aspect of plant rooting strategies to access underground resources.*

Keywords: Meristem dynamics, development, architecture, wave, root-soil interaction.

1 Introduction

Land plants grow in soil where water and nutrients are heterogeneously distributed. Growth, reproductive success, and chance of survival are largely conditioned by the plant ability to acquire these resources efficiently ([2]). There is considerable evidence for the influence of root architecture on water and nutrient acquisition efficiency. However, fundamental mechanisms by which root architecture develops and adapts to environmental conditions are complex and poorly understood. Root architectures result from the activity of their meristems. Meristems develop in a sequence of expansion and lateral initiation events at the proximity of root apices, but the detailed mechanisms by which water and nutrients are sensed by plants remain idle. There is increasing evidence that the sensing activity in roots, first postulated by ([1]) is concentrated in apical meristems. Unfortunately much less is known about the way meristems proliferate in soil. This is due notably to the difficulty to experiment in soil without perturbing the growth of the plant. There has been considerable effort to develop non-destructive root imaging methods, for example Xray tomography, rhizotrons and gel observation chambers ([6]), but growth conditions in such experimental systems barely represent those found in the field. Root architectural models have provided great insight into root developmental processes ([3]). However, estimating parameters for architectural models has always suffered from

the requirement for time consuming, tedious and destructive experiments for which the final accuracy of the spatial data is often limited.

In this study, we have developed a mathematical framework to study the dynamical behaviour of root meristems. We have set up a minirhizotron experiment to collect data on root length distribution and we used the relationship between changes in root length density and meristem density to identify regions of meristematic activity in soil. Finally, we have constructed a mathematical model of root meristems dynamics and this was used to understand the patterns observed with the experimental data. A *supplementary website* (`http://www.scri.ac.uk/research/epi/resourcecapture/plantmodelling/meristemwave`) offers access to the code (Python) used in this work, to additional experiments results and to technical annexes.

Material and method

Barley (Hordeum vulgare L.) cv. Optic was grown in soil in two concrete bins (4 rows) in a glasshouse, with drainage in the base, filled with soil (of known characteristics see supplementary website) in layers and repeatedly irrigated prior to sowing. Clear perspex minirhizotron access tubes (3 in each bin at 6 different depths) were placed horizontally across the width of the concrete bin. Irrigation was applied to replace evaporation. A minirhizotron camera was inserted every day into each tube and images were captured showing root impacts on the tube ($1.25m$ long, $50mm$ diameter) on $1.2cm \times 0.8cm$ images (suitable resolution to distinguish living roots accurately). The experiment was started 8 days after sowing once few root impacts were observed in the shallowest tube, and was stopped 9 days later, when root impacts started to be recorded on the deepest tubes. Individual root lengths were measured manually from digital images using in-house software. The depth z and distance x from the row are defined respectively from the depth of the tube and the position of the camera within the tube. α denotes the angle of the root with the horizontal direction. Root length density curves were derived from the number measurements of root lengths using a mean filter with a window of $5cm$.

2 Theoretical framework

2.1 Analysis of root meristem dynamics

Plant root systems can be characterized using density distributions. For instance, root length density, ρ_n (cm^{-1}) and root branching density, ρ_b (cm^{-2}) describe the architecture of roots ([7]). Root meristem density ρ_a (number of meristems per length area or volume) indicates regions of primary growth and defines the sensing compartment of the root system. Densities are multivariate functions depending upon spatial position r, time, and the incline angle of the roots.

In general, it is not feasible to track individual roots and their meristems in soil. However, the architecture of the root system at time t results from the functioning of all meristems during growth. Relationships can be derived to obtain meristem functioning properties:

$$\frac{\partial \rho_b}{\partial t} = \quad b \qquad \text{branching rate } (cm^{-3}d^{-1})$$

$$\frac{\partial \rho_n}{\partial t} = \rho_a.e \qquad \text{meristem activity } (cm^{-2}d^{-1})$$

The meristem activity can be determined directly from time sequences of the root length density profiles. We analyzed temporal patterns of the meristem activity by combining two different indicators $I_a(z,t) = \int_x \int_\alpha \dot{\rho_n}(x,z,\alpha,t)dx$ (change in root length $cm.d^{-1}$) and $I_g = I_a(x_g - x_c)$ (position of meristem distribution $cm^2.d^{-1}$, x_g centre of mass of $\dot{\rho_n}$, x_c geometric centre of the distribution of roots). I_a is proportional to the area between two consecutive root length density curves hence measures the intensity of the root meristem activity at a given time. I_g is a measure of the centre of mass of the meristem density along the x axis. Because the size of the rooting domain grows with time, we chose a measure of the centre of mass relative to x_c in the domain to describe the bulk position of meristems. Plots representing the trajectories in the (I_a, I_g) space from both experimental data (Fig. 1 C) and model predictions (Fig. 1 D) were used to analyze the mechanisms of root proliferation in soil.

2.2 Modelling the dynamics of root growth

To understand the experimentally observed dynamic patterns of root meristematic activity, we have built a simple mechanistic model that describes how meristem distribution evolves as a function of the root expansion rate, gravitropism and branching rate. The model is derived by a continuity equation:

$$\frac{\partial \rho_a}{\partial t} + \nabla^*(\rho_a g) + \nabla(\rho_a e u) = b, \text{ (where } u = (cos(\alpha), sin(\alpha)) \text{ is the root growth direction)}$$

∇ and ∇^* are gradient operators for the spatial coordinates and the direction of growth, respectively. In this equation, the change of root meristem density results from the number of roots entering the neighbourhood from upstream regions or leaving the neighbourhood at expansion rate e ($cm.d^{-1}$), the number of roots changing their orientation through gravitropism g (d^{-1}) and the creation of new meristems through branching rate b ($cm^{-2}.d^{-1}$). In this model and throughout this study, we ignore root mortality, although it could be incorporated in the model as a sink term in b. This simplification is justified as we considered early stage growth, and no root decay was observed in rhizotron images.

This model can be adapted to account for different root behaviours based on their branching order (number of connections required to link a root to the stem): $\rho_{1,a}$ describes the meristem density of first branching order roots and $\rho_{2,a}$ the density of second branching order roots having different growth rates and gravitropism:

$$\frac{\partial \rho_{1/2,a}}{\partial t} + \frac{\partial \rho_{1/2,a} g_{1/2}}{\partial \alpha} + \frac{\partial \rho_{1/2,a} e_{1/2} \cos(\alpha)}{\partial x} + \frac{\partial \rho_{1/2,a} e_{1/2} \sin(\alpha)}{\partial z} = 0/b_2$$

The two equations are coupled through the source term b_2 which is a function of ρ_1. Terms e, g and b are not constants in general but functions that encode both developmental behaviour and root/soil interactions. To simplify the analysis of the model, we used arbitrary functions that illustrate the type of mechanisms observed on real systems: $e_1 = e_2 s = cst$, $g_1 = g_2 s^2 = g_{11}(\pi/2 - \alpha)$ and $b_2 = b_{21}(\rho_{1,a}(r, u + b_{22}, t) + \rho_{1,a}(r, u - b_{22}, t))/2 + b_{23}\rho_{2,n}$, where s is a scaling factor, e_i's are constant root expansion rates, g_i's are root vertical orientation rates, b_{21} (d^{-1}) is the branching rate, b_{22} (d^{-1}) is the branching angle and b_{23} ($cm^{-2}.d^{-1}$) is the adventitious branching rate (lateral root initiation on mature tissues).

A one dimensional solution has been developed to obtain root developmental parameters (g, e, b) and assess how the model can predict experimental root distributions:

$$\rho(z,.,t) = a(1+bt)\exp\left(-(z - et - e\alpha^2/4g(e^{-2gt}-1))^2/c\right)$$

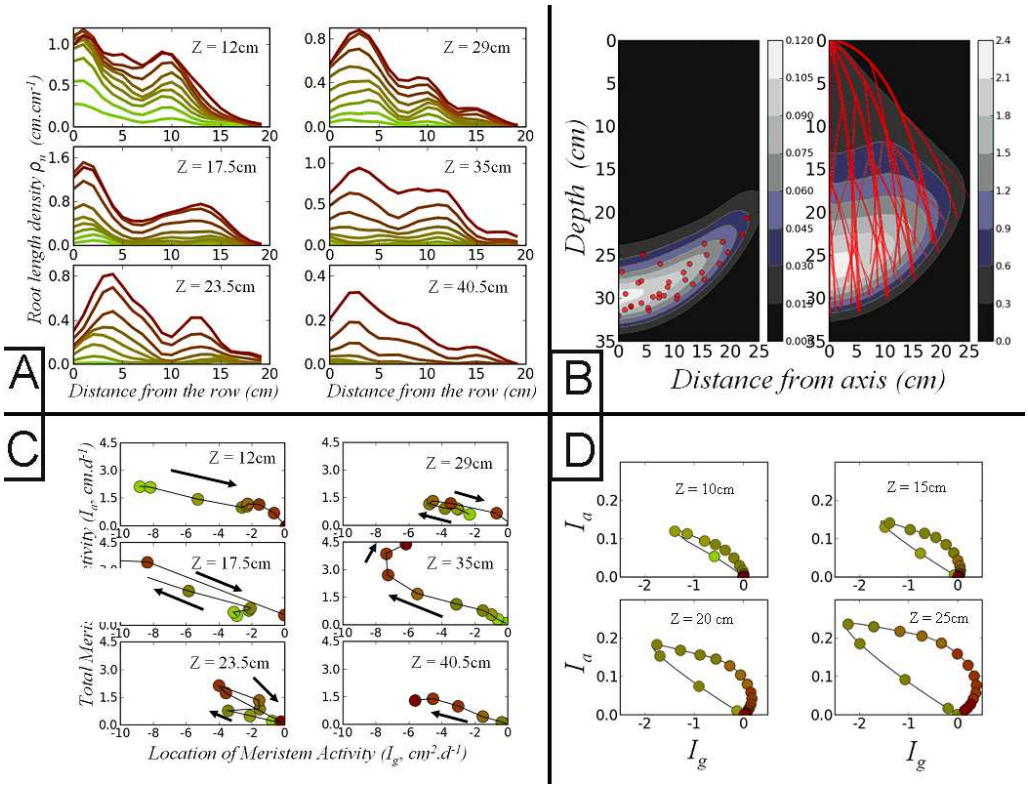


Figure 1. A: Patterns of Barley root length density distribution at different depths (gray levels represent measurement days) vs distance from the row x. B: Root length (left) and root meristem (right) densities (contour plots) compared with meristem positions (circles) or whole root system (lines) for numerical simulations. C: Barley meristem activity in soil during growth at different depths in the (I_g, I_a) (cf. text in 2.1) space. D: Trajectories at different depths of the meristem position in soil using indicators I_a and I_g (cf. text in 2.1) for numerical simulations.

2.3 Numerical analysis

We have developed a numerical solver to explore the dynamic patterns generated by meristematic activity in soil as described in 2.2. The equation can be solved numerically using a finite volume method ([4]). We used a high resolution upwind scheme (minmod) and dimensional splitting to compute root fluxes on the domain. The root zone consists in a rectangular domain of $30cm \times 40cm \times \pi$ divided into $30 \times 40 \times 40 = 48,000$ control volumes to account for the spatial coordinates and the root angle. The simulation is initiated in the top left corner of the grid by a uniform meristem density of 1 for $0 \leq \alpha \leq \pi$. At the boundaries of the domain, symmetric fluxes are imposed on the vertical plane and no root flux is permitted on the remaining planes.

Although the model presented in 2.2 represents root systems as density functions, it uses fundamental developmental parameters as used in classical root architectural models. We have checked on numerical solutions by comparing a simulated root architecture (Fig. 1 B) from an equivalent model

with the results of the numerical solution of the model. Looking at the one dimensional (depth) analytical model, we noted that meristems form a peak of activity which propagates downwards. Comparing model predictions and measurements of root meristem activity as a function of depth, we found a good agreement between predicted and measured meristem activity ($R^2 = 0.85$, unpublished data) with a reasonable number of parameters (4 compared to the 9 measured times $\times$ 6 tubes). The behaviour of the model was then analyzed on a range of input parameters. 15 simulations were run, with each parameter taking successively low, medium and high values (other parameters fixed at their medium values). Meristem density distribution and root length density distribution were visualized using contour plots (see Fig. 1 B).

3 Results

3.1 Minirhizotron data indicates that root meristem distribution has spatial and temporal patterns

Root length density profiles were determined for each minirhizotron tube on the 9 days of measurements (Fig. 1 A). Some interesting features were observed: (i) first root impacts appear at the proximity of the row of barley crops, (ii) root length density increases with time and propagates gradually at larger distances from the row, (iii) root length density peaks where first roots initially appeared and (iv) the maximum of the root length density is gradually shifted away from the row when time increases. Such meristem proliferation process was captured by visualizing the position of the bulk of the meristem activity using both indicators I_a (total meristem activity) and I_g (centre of gravity of meristem activity). During growth, the profile of the meristematic activity takes the form of trajectories in the (I_a, I_g) space (Fig. 1 C). It expresses what was observed qualitatively on root length density profiles: meristems first hit tubes at the proximity of the row. Therefore, the centre of mass of the meristematic activity is shifted to the left. In a second stage, a peak is reached and meristematic activity is shifted from the row while decreasing. Finally, most meristem activity disappears and the trajectory returns to its initial position. These preliminary results indicate a meristematic activity front that propagates wave-like during growth. The velocity of propagation determines when meristem activity first reaches a given depth and then disappears deeper in the soil.

3.2 Models show that meristems propagate like waves

Using the fundamental principle of conservation, we derived a general (continuity) equation relating root growth processes and distribution of meristems in the soil (section 2.2). This equation is categorized as a hyperbolic partial differential equation, which occurs frequently in the study of wave phenomena (acoustic, mechanics, electromagnetism) in physics. This indicates that meristematic waves may form in the soil, as a result of root developmental processes. The analytical solution of the simple model developed at the end of 2.2 is in good agreement with experimental data (unpublished data) in addition to providing interpretable developmental parameters. Numerical analysis of the model consisted of 15 days of growth inititiated punctually at the surface of the domain (Fig. 1 B and D). During growth, these regions gradually expand and progress downwards. Root length density distribution is determined as the accumulation of meristem production during the simulation, and can be seen as the footprint of the meristematic activity. Predictions from our model were compared with the simple architectural model ([7]). It was found that both root length density distribution and root meristem distribution were visually in good agreement (Fig. 1 B). The trajectories of the meristem distributions in the (I_a, I_g) space showed similar features (Fig. 1 D), providing additional evidence

for a wave propagation mechanism underlying the experimental observations. We were also able to analyze dynamic patterns of root development in soil by simulating the influence of gravitropism, branching rate or angle, root expansion rate and adventious branching (results not shown).

4 Discussion/Conclusion

Plant architectural models have greatly contributed to our understanding of plant growth and interaction with the environment. Classical models describe the deployment of the plant architecture simulating the behaviour of each organ at each time step. While convenient to dissect developmental processes, it has also obvious limitations: computer time when the number of roots becomes high, interpretation of the system property using a 3D architectural models (a priori more detailed) and need for accurate measurements of single organ properties while suffering from a difficult parametrization, and resulting simulations are prone to error propagation. Our work uses a different approach to architectural modelling whereby root system development is represented through spatial density distributions (see [5] for earlier studies fitting a diffusion model). [8] developed a similar model focusing on nutrient intake but didn't fully addressed root architecture. Moreover our choice to work in a deterministic framework allowed us to explicitly represent the global dynamics. We provided a new mathematical framework for quantitative plant architectural analysis. It has overcome the shortcomings of reaction-diffusion based models by incorporating architectural parameters explicitly. Our approach could be particularly useful to implement plant growth processes at larger scales of application. Our model could also be beneficial to study rooting strategies for optimized resources acquisition by incorporating the knowledge that meristematic wave-like activity and its localization in specific regions of the soil. A particularly appealing development we would like to address in a further study for our model would be the integration of physiological and molecular data, still a great challenge at an entire plant level. The challenge is now to extend these concepts in order to integrate plant/environment dynamical feedbacks.

Acknowledgements

We would like to thank Ian McNaughton, Jim Anderson, Jacqueline Thompson, Tracy Valentine and Glyn Bengough for help and comments on the present work. The SCRI receives grant-in-aid support from the Scottish Government RERAD (Workpackage 1.7).

References

[1] MC. Drew, Comparison of the effects of a localised supply of phosphate, nitrate, amonium and potassium in the growth of seminal root system, and the shoot in barley. *New Phytol*, 75:479-490, 1975.
[2] R. Aerts et al., The relation between above and belowground biomass allocation patterns and competitive ability. *Oecologia* 87:551-559, 1991.
[3] C. Jourdan and H. Rey, Modelling and simulation of the architecture and development of the oil-palm (Elaeis guineensis jacq.) root system. *Plant Soil*, 190:217-233, 1997.
[4] RJ. Leveque, *Finite volume methods for hyperbolic problems*, Cambridge University Press, 2002.
[5] M. Heinen et al., Growth of a root system described as a diffusion: 2 numerical models and application. *Plant Soil*, 252:251-265, 2003.
[6] PJ. Gregory et al., Non-invasive imaging of roots with high resolution x-ray micro-tomography. *Plant Soil*, 255: 351-359, 2003.
[7] L. Dupuy, et al., A density-based approach for the modelling of root architecture: application to maritime pine (pinus pinaster ait.) root systems. *J Theor Biol*, 236:323-334, 2005.
[8] P. Bastian et al., Modelling in vitro growth of dense root networks. *J Theor Biol*, 254:99-109, 2008.

Construction et analyse d'un modèle tridimensionnel du complexe [(SLR1738-Zn-Fe)$_2$-ADN]

Paul Garcin[1], Olivier Delalande[1], Corinne Cassier-Chauvat[1], Franck Chauvat[1] et Yves Boulard[1]

[1] Laboratoire de Biologie Intégrative, CEA Saclay, 91191 Gif-sur-Yvette, France
{paul.garcin, corinne.cassier-chauvat, franck.chauvat, yves.boulard}@cea.fr
olivier.delalande@ibpc.fr

Abstract: *Slr1738 is the Peroxide regulon Repressor protein (PerR) of the cyanobacteria Synechocystis. Active as a dimer, this protein must contain an iron atom to be able to bind DNA molecule and regulates targeted genes. The binding mechanism involves a classic recognition helix inserted in the DNA major groove. But to date there is no three-dimensional structure available for this kind of transcription factor complexed to DNA. As a consequence, both global and specific interactions that lead this protein to bind DNA and to recognize specific 'Per Box' sequence are still misunderstood. In order to better define and analyse these interactions, we built in silico the first three-dimensional structure of a [PerR-DNA] complex. This article describes the method used to build the complex and presents an analysis of the contacts between the two partners.*

Keywords: PerR, transcription factor, protein-DNA complex, HTH recognition motif, modelisation.

1 Introduction

Au sein de tous les organismes vivants, la régulation des dérivés réactifs de l'oxygène (ROS) est essentielle à la viabilité des cellules. Leurs effets nocifs sur les processus cellulaires ont été largement décrits dans la littérature [1]. L'interaction des ROS avec différentes molécules biologiques, notamment les acides nucléiques, peut induire l'apoptose cellulaire. Ainsi, un excès de ROS conduit à un état de stress oxydatif qui doit être contrôlé.

Les protéines FUR (Ferric Uptake Repressor) sont connues pour leur implication dans l'homéostasie des métaux, et notamment du fer [2]. Cette grande famille de métallo-régulateurs regroupe plusieurs types de répresseurs dont Fur, Zur (Zinc uptake repressor) et PerR (Peroxide regulon Repressor) en sont les principaux représentants. Cette dernière sous-famille joue un rôle important comme senseur de stress oxydatif et est donc étroitement liée aux mécanismes cellulaires impliqués dans la gestion des dérivés réactifs de l'oxygène.

Dans ce cadre nous étudions le facteur de transcription Slr1738, une protéine FUR de la cyanobactérie *Synechocystis*. Une étude récente a montré que Slr1738 est impliquée dans la résistance de cette cyanobactérie à un stress oxydant [3]. Il a également été démontré que le mutant ΔSlr1738 est plus résistant que la souche sauvage à un stress oxydant provoqué par H_2O_2 mais aussi plus réactif à un stress métallique (Cd, U et Se). Ces résultats suggèrent que

Slr1738 serait la PerR de *Synechocystis*.

À ce jour, il n'existe pas de structure tridimensionnelle de la famille des facteurs de transcription FUR complexé à l'ADN. Par ailleurs, aucun programme de modélisation ne permet, contrairement au cas des protéines, de construire un modèle 3D d'un complexe [protéine-ADN]. C'est pourquoi, dans le but d'étudier en détail les propriétés de la protéine Slr1738 et les interactions qu'elle établit avec l'ADN, nous avons développé une méthode permettant à partir de la structure primaire de Slr1738 d'élaborer *in silico* le premier modèle d'un complexe [PerR-ADN].

2 Matériels et méthodes

2.1 Mécanique moléculaire et dynamique moléculaire

Les simulations ont été réalisées à l'aide de la suite de logiciels contenu dans AMBER 9 et en utilisant le champ de force Parm99. Chaque système étudié est neutralisé (ajout d'ions Na^+) et hydraté avec des molécules d'eau du type TIP/3P avant minimisation. La minimisation est réalisée en six étape. Pendant les 5 premières, on relâche progressivement (de 100 à 5kcal/mol) les contraintes appliquées sur le soluté. La dernière étape réalisée sans contrainte aboutit à l'obtention d'une structure stable d'un point de vue énergétique.

La simulation de dynamique moléculaire a été réalisée à 300K dans l'ensemble NVT, à pression et température constantes. La méthode « Particle Mesh Ewald » est utilisée pour le calcul des interactions électrostatiques, avec un cutoff dans l'espace direct fixé à 9 angströms.

2.2 MM-(GB)PBSA

La méthode MM-PBSA [4] (Molecular Mechanics Poisson-Boltzman Surface Area) a été utilisée afin de calculer l'énergie libre d'association entre un récepteur et son ligand. La variation d'énergie libre est définie selon l'équation ci-dessous.

$$\Delta G = G_{Complexe} - G_{\text{Pr}\,otéine} - G_{ADN}$$

2.2 MSMS

La surface de contact entre la protéine et l'ADN d'un complexe a été évaluée en utilisant le programme MSMS. Ce dernier calcule la valeur de la surface accessible au solvant (SAS) d'une structure. La valeur de la surface de contact entre un récepteur et son ligand est alors obtenue à l'aide de la formule ci-après :

$$SC = \frac{SAS_{rec} + SAS_{lig} - SAS_{cplx}}{2},$$

où SAS_{rec}, SAS_{lig} et SAS_{cplx} sont respectivement les valeurs de surface accessible au solvant du récepteur, du ligand et du complexe.

3 Résultats

3.1 Construction du monomère – Slr1738

La structure du monomère de Slr1738 (cf. figure 1) a été construite par homologie à partir de la protéine Fur de *Pseudomonas aeruginosa* (code PDB: 1MZB [5]). Cette dernière possède 21% d'identité et 37% d'homologie avec Slr1738. Les protéines FUR sont constituées de deux domaines fonctionnels distincts. Il y a un domaine N-terminal, dit domaine de liaison à l'ADN,

et un domaine C-terminal impliqué dans le processus de dimérisation de la protéine. Pour Slr1738, le domaine N-terminal (résidus 1 à 84) est composé de quatre hélices α (H1-H4) suivies de deux brins β (F1, F2) formant un feuillet β anti-parallèle. Le motif wHTH (winged-Helix-Turn-Helix), qui permet l'interaction avec l'ADN, est constitué de H3, H4, F1 et F2, l'hélice H4 étant l'hélice de reconnaissance. Le domaine C-terminal (résidus 85 à 139) est composé de deux brins β (F3, F4), d'une hélice α (H5) et d'un dernier brin β (F5).

3.2 Les Sites métalliques – Slr1738-Zn-Fe

Slr1738 contient deux sites métalliques. Il y a un site à atome de zinc, dit site structural (1), impliqué dans la dimérisation de la protéine et un site régulateur (2) capable de lier un atome de fer afin d'activer la protéine et lui permettre sa liaison à l'ADN. La localisation des deux sites est respectivement représentée par des sphères de van der Waals sur la figure 1 ci-dessous.

Le site zinc est composé de quatre ligands cystéines ($Zn(Cys)_4$) et se présente sous la forme d'une sphère de coordination tétraédrique. Des paramètres de champ de forces pour ce site avaient déjà été proposés par les équipes de Ryde *et al*. et Stote *et al*. [6, 7]. Durant les simulations, l'ion Zn^{2+} est lié de façon covalente aux 4 cystéines C95, C98, C134 et C137.

Le site fer, qui est impliqué dans le rôle de senseur de stress oxydant (H_2O_2) [8, 9], est moins bien décrit que le site zinc dans la littérature. Pour réaliser une paramétrisation de ce site, nous avons utilisé les données de McLuskey *et al*. et de Sundar [10, 11]. Son environnement est composé de quatre à cinq ligands potentiels (H36, D84, H90, H92 et D103). Ce site peut alors être penta- ou hexa-coordonné sous la forme d'une bi-pyramide à base carrée. Des contraintes harmoniques de distance ont été appliquées entre l'ion et ses ligands durant les simulations afin de maintenir son intégrité.

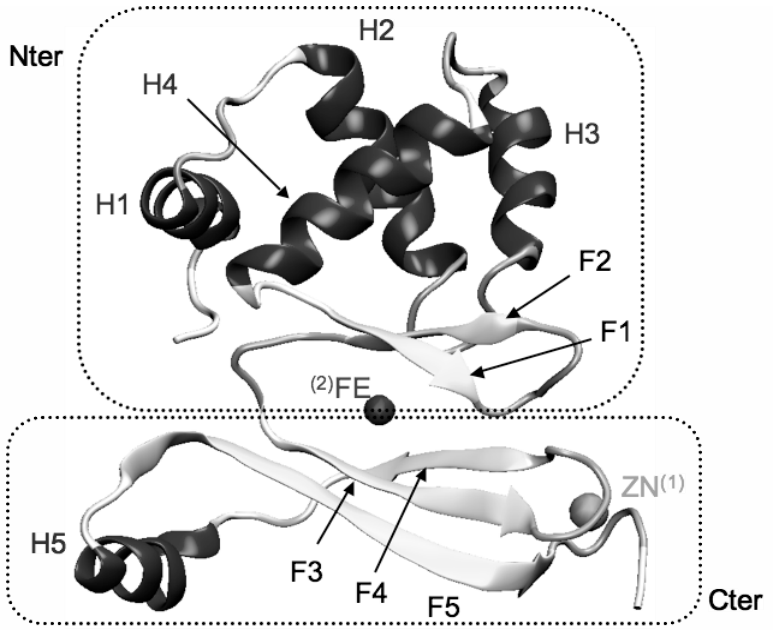


Figure 1. Représentation du monomère (Slr1738-Zn-Fe).

3.3 Construction du dimère – $(Slr1738\text{-}Zn\text{-}Fe)_2$

Nous avons ensuite construit le dimère de Slr1738, qui correspond à la forme active de la protéine, par arrimage de deux monomères. Ce travail a été réalisé en deux temps. Nous avons d'abord créé un premier modèle, la structure 1MZB (Fur de *Pseudomonas aeruginosa*) nous

ayant servi de référence [5]. Puis nous avons utilisé la structure d'une PerR de *Bacillus subtilis* (code PDB : 2FE3 [12]) publiée durant notre étude. Cette structure présentant un dimère inactif, car non chargé en fer, nous a été très utile afin d'optimiser l'interface de dimérisation du modèle. Celle-ci implique le domaine C-terminal de chacun des deux monomères A et B, et est principalement constituée d'un feuillet β anti-parallèle entre les brins $F5_A$ et $F5_B$. On notera la présence de nombreux résidus hydrophobes au sein de cette interface ainsi que la formation de nombreux ponts salins favorisant l'interaction des deux monomères.

3.4 Construction du complexe – [(SLR1738-Zn-Fe)$_2$-ADN]

L'objectif de notre travail étant de construire un modèle tridimensionnel fiable du complexe [(SLR1738-Zn-Fe)$_2$-ADN], nous avons mis en place un protocole de modélisation d'abord basé sur des informations structurales et topologiques tirées de l'analyse de structures de complexes expérimentaux disponibles dans les bases de données.

La sélection des structures expérimentales repose sur trois critères : 1) la qualité de l'homologie avec le domaine de liaison à l'ADN de Slr1738, à savoir le motif de reconnaissance wHTH (longueur du motif et nature des résidus), 2) la taille du fragment d'ADN et 3) la résolution des structures expérimentales. À l'aide de ces critères, nous avons finalement sélectionné 4 structures sur les 1214 complexes existants. Elles proviennent d'organismes variés et aucune n'appartient à la famille des FUR. Ces quatre structures, 1C0W, 1SAX, 1U8R et 1Z9C, nous ont permis de générer plusieurs modèles structuraux initiaux afin d'explorer différents types de complexation.

La première étape de construction a consisté à positionner un monomère de Slr1738 sur l'ADN des structures expérimentales sélectionnées en le superposant au monomère de la protéine du complexe sélectionné. Dans le but d'obtenir un bon positionnement du motif de reconnaissance de Slr1738 par rapport à l'ADN, nous avons réalisé quatre superpositions différentes basées sur la structure secondaire du motif de reconnaissance wHTH : H4, H3-H4, H4-F1-F2 et H3-H4-F1-F2. Après minimisation des systèmes, la surface de contact est évaluée pour chaque complexe. Les meilleurs résultats ont été obtenus avec les superpositions H4-F1-F2 et H3-H4-F1-F2 et 8 modèles ont ainsi été retenus. Ensuite, pour ces 8 structures nous avons fixé la taille des fragments d'ADN cristallographiques à 25 paires de bases, taille minimale correspondant à l'interaction avec Slr1738. La séquence d'ADN de la structure cristallographique est alors substituée par la séquence correspond à la région intergénique sur laquelle Slr1738 vient se fixer. Finalement une sélection basée à la fois sur le calcul de l'énergie d'association et la surface de contact nous a permis de retenir 3 modèles (reportés en gras dans le tableau 1) sur les 8 pré-sélectionnés.

	H4F1F2		H3H4F1F2	
1C0W	35kcal	900Å^2	19kcal	888Å^2
1SAX	24kcal	889Å^2	**13kcal**	**946Å^2**
1U8R	26kcal	949Å^2	**7kcal**	**885Å^2**
1Z9C	**-7kcal**	**1006Å^2**	11kcal	938Å^2

Table 1. Tableau de comparaison pour les différents complexes construits à l'étape 1.

À ce stade, les structures construites ne possèdent qu'un seul monomère correctement arrimé à l'ADN. La deuxième étape a donc consisté à placer correctement le second monomère sans altérer la position du premier. Pour y parvenir, différents protocoles de minimisations sous contraintes harmoniques de distances ont été testés en faisant varier le nombre et la nature des contraintes appliquées (approche du type fragment rigide). Le protocole le plus performant et le moins déstructurant fut utilisé sur les 3 modèles retenus précédemment. Finalement, la structure qui donne les meilleurs résultats est celle qui a été obtenue à partir des informations topologiques de la structure 1SAX [13] (avec la superposition H3-H4-S1-S2). Par cette méthode, nous avons réussi à construire un modèle de Slr1738 compléxé à un double brin d'ADN de 25 paires de bases.

3.5 Validation du complexe

Afin de valider le modèle construit, nous avons effectué une simulation de dynamique moléculaire sur le complexe [(SLR1738-Zn-Fe)$_2$-ADN] sans fixer de contrainte au niveau de l'interface protéine-ADN. Le système comporte environ 6000 atomes de soluté dans une boîte d'environ 15000 molécules d'eau. Le complexe reste stable pendant l'ensemble du calcul. On peut également noter que les hélices de reconnaissance restent insérées dans le grand sillon, que la surface de contact du complexe de même que l'énergie du système restent constantes.

3.6 Analyse des contacts protéine-ADN

Le modèle retenu présente trois zones de contact entre l'ADN et chaque monomère de la protéine (cf. figure 2). La première zone (a) implique les résidus qui se situent au niveau du coude entre H1 et H2. La seconde (b) est composée de l'hélice de reconnaissance H4 et de quelques résidus du coude en amont de cette hélice. La dernière zone (c) regroupe les résidus entre les brins F1 et F2.

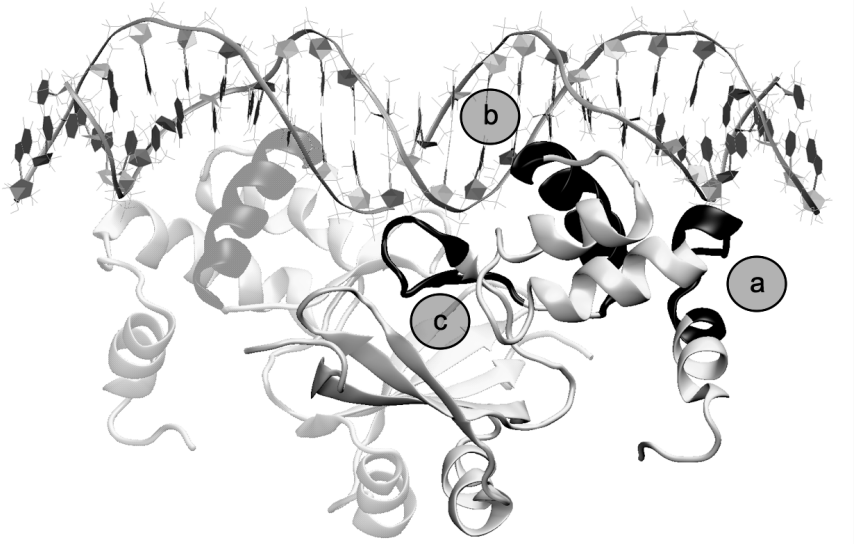


Figure 2. Représentation du modèle du complexe [(SLR1738-Zn-Fe)$_2$-ADN]. Les zones de contact entre la protéine et l'ADN (a, b, c) sont en noir.

(a). Cette région de la protéine est composée de cinq acides aminés chargés positivement (^{13}**K**E**R**GL**R**VTP**QR**^{23}). Trois d'entre eux, R18, Q22 et R23, sont impliqués dans des interactions électrostatiques avec les charges négatives des groupements phosphates du petit sillon de l'ADN. Des alignements de séquence avec d'autres PerR montrent que les résidus K13, R18, Q22 et R23 sont très conservés. Nous considérons cette première zone d'interaction comme non spécifique dans la mesure où il n'y a pas de contact entre les chaînes latérales de ces acides aminés et les bases de l'ADN. Nous pouvons supposer que ces interactions électrostatiques longues portées peuvent participer à une pré-orientation initiale du domaine N-terminal et à une stabilisation du complexe après la liaison à l'ADN.

(b). L'hélice de reconnaissance H4 de Slr1738 est composée de 16 acides aminés (^{54}SQATVYSSLKALQSVG^{69}). Une description détaillée de cette hélice montre qu'elle peut être sub-divisée en « trois faces ». La première regroupe des résidus hydrophobes (V58, L62 et L65). Ils sont en contact direct avec d'autres acides aminés hydrophobes présents dans les trois premières hélices du domaine N-terminal et forment ainsi un cluster hydrophobe compact. La seconde face est composée de sept petit résidus (A56, T57, S60, S61, A64, S67 et V68) qui sont proches d'un des brins de l'ADN. Enfin, des acides aminés encombrants et chargés (Q55, Y59, K63 et Q66) composent la dernière face orientée vers le second brin d'ADN. Les résidus au

contact de la molécule d'ADN, et notamment des bases, appartiennent aux faces 2 et 3 de l'hélice H4 : Q55, A56, T57, Y59, S60 et K63. L'alignement de protéines FUR montre que le motif TVY présent dans l'hélice de reconnaissance est toujours très conservé. D'autre part, pour Slr1738 on retrouve une grande concentration de résidus hydroxyles (5 sérines et 2 thréonines) qui semblent être importants pour la spécificité de reconnaissance des 'PerR Box' de *Synechocystis*.

(c). Malgré sa proximité, la dernière région au contact de l'ADN (^{73}**E**VLLE**E**GVC^{81}) ne montre pas d'interaction favorable entre la protéine et l'ADN. Au cours de la DM, on constate d'ailleurs un léger mouvement d'éloignement du feuillet composé par F1 et F2 dû à la présence des acides glutamiques E73 et E78. Nous pensons que ce mouvement pourrait induire un meilleur positionnement général du domaine N-terminal de liaison à l'ADN.

4 Conclusion

Dans cet article, nous présentons une méthode utilisée pour la construction d'un modèle tridimensionnel du facteur de transcription Slr1738 compléxé à une molécule d'ADN. Pour valider ce complexe, nous avons effectué une dynamique moléculaire qui démontre la bonne stabilité du modèle. Une première analyse montre l'existence de trois zones de contact entre la protéine et l'ADN. La première zone (a) permet clairement de stabiliser le complexe protéine-ADN par la formation d'interactions non spécifiques. Le rôle de la zone (c) doit servir à optimiser l'orientation du domaine N-terminal. Finalement, la zone (b), principalement composé par l'hélice de reconnaissance, montre de nombreux contacts électrostatiques et hydrophobes avec l'ADN. Seule zone à former des interactions directes avec les bases de l'ADN, elle est responsable de la spécificité de reconnaissance. L'identification d'interactions non spécifiques montre qu'elle contribue également à la stabilisation du complexe.

Références

[1] Storz et Imlay, Oxidative stress. Curr. Opin. Microbiol., 2:188-194, 1999.
[2] Hantke, Iron and metal regulation in bacteria. Curr. Opin. Microbiol., 4: 172, 2001
[3] Houot et al., Cadmium triggers an integrated reprogramming of the metabolism of Synechocystis PCC6803, under the control of the Slr1738 regulator. BMC Genomics 8: 350, 2007
[4] Kollman et al., Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. Acc. Chem. Res. 33:889-97, 2000
[5] Pohl et al., Architecture of a protein central to iron homeostasis: crystal structure and spectroscopic analysis of the ferric uptake regulator. Mol. Microbiol., 47(4):903-15, 2003
[6] Ryde et al., Molecular dynamics simulations of alcohol dehydrogenase with a four- or five-coordinate catalytic zinc ion. Proteins, 21(1):40-56, 1995
[7] Stote et al., Zinc binding in proteins and solution: a simple but accurate nonbonded representation. Proteins, 23(1):12-31, 1995
[8] Lee et Helmann, The PerRtranscription factor senses H_2O_2 by metal-catalysed histidine oxydation. Nature, 400: 363-367, 2006
[9] Traoré et al., Structural and functional characterisation of 2-oxo-histidine in oxidized PerR protein. Nat. Chem. Biol., 5(1):53-9, 2009
[10] McLuskey et al., Structure and reactivity of hydroxypropylphosphonic acid epoxidase in fosfomycin biosynthesis by a cation- and flavin-dependent mechanism. PNAS, 102(40):14221-6, 2005
[11] Sundar. Thesis. 1997
[12] D. A. K. Traoré et al. Crystal structure of the apo-PerE-Zn protein from Bacillus subtilis. Mol. Microbiol., 61(5):1211-9, 2006
[13] Wisedchaisri et al. Crystal Structure of an IdeR-DNA Complex Reveals a Conformational Change in Activated IdeR for Base-specific Interactions. J. Mol. Biol., 342: 1155-1169, 2004

A geometric knowledge-based coarse-grained scoring potential for structure prediction evaluation

Sébastien Loriot[1], Frédéric Cazals[1], Michael Levitt[2], Julie Bernauer[1,2] ⋆

[1] Algorithms, Biology, Structure project-team, INRIA Sophia Antipolis
2004 route des Lucioles, BP 93, 06902 Sophia-Antipolis, France
firstname.lastname@sophia.inria.fr
[2] Department of Structural Biology, Stanford University School of Medicine
Stanford, CA 94305-5126, USA
michael.levitt@stanford.edu

Abstract: *Knowledge-based protein folding potentials have proven successful in the recent years. Based on statistics of observed interatomic distances, they generally encode pairwise contact information. In this study we present a method that derives multi-body contact potentials from measurements of surface areas using coarse-grained protein models. The measurements are made using a newly implemented geometric construction: the arrangement of circles on a sphere. This construction allows the definition of residue covering areas which are used as parameters to build functions able to distinguish native structures from decoys. These functions, encoding up to 5-body contacts are evaluated on a reference set of 66 structures and its 45000 decoys, and also on the often used* lattice_ssfit *set from the* decoys'R *us database. We show that the most relevant information for discrimination resides in 2- and 3-body contacts. The potentials we have obtained can be used for evaluation of putative structural models; they could also lead to different types of structure refinement techniques that use multi-body interactions.*

Keywords: knowledge-based potential, structure prediction and refinement, spherical arrangements, surface area, coarse-grained model.

1 Introduction

Amongst the forces driving protein folding, solvent effects and hydrophobic interactions are known to play the greatest role [2]. Calculation of the solvent accessible surface area has given important insights to estimate solvation energies [10,13]: methods that estimate solvation energies from surface area have proven useful for physics-based potentials [8,9]. Knowledge-based potentials, built from structures that are known to be stable in solution are expected to take solvation effects into account. These potentials are generally derived from distance measurements in known protein structures. For example, comparing the distribution of distances between two hydrophobic residues and that between a hydrophobic residue and a hydrophilic residue, shows that hydrophobic residues minimize their solvent contact.
The theoretical basis of such knowledge-based potentials have been questioned [3] but they often have proven to be as successful as physics-based potentials [12,15,19,22,24,25]. Demonstrating the validity of knowledge-based potentials has become easier with the availability of large, and good-quality

⋆ The authors thank the France-Stanford Center for Interdisciplinary Studies and the INRIA Équipes Associées program for funding, and the NSF award CNS-0619926 for computer resources.

protein decoy datasets [20,25], partly triggered by the CASP experiment (http://predictioncenter.org/). Attempts to derive knowledge-based potentials from more precise definitions, such as the Voronoi tessellation procedure, are as efficient as distance-based techniques [17]. The contacts obtained using this type of procedure often give a sharper signal, providing a more accurate description that leads to a better performing potential or scoring function [4].
Although fast and accurate accessible surface area calculations [1,5] and two- and four-body potentials [11,18,16,21] have been developed, none has addressed the problem of multi-body contact area. Recent improvements in computations of spherical arrangements give quick access to the detailed buried areas of a sphere intersected by other spheres. This has allowed us to derive potentials from these surface area computations, and consider different terms that range from accessible surface area to multi-body contacts. We show that these potentials that use an accurate description of local environments, provide a good discrimination between native/near-native structures and decoys.

2 Material and Methods

Geometric Construction. Given a set of $n+1$ spheres $S_{i,i=0\ldots n}$ in 3-dimensional space, we consider a tuple of $k+1$ pairwise intersecting spheres $S_{i_0}\ldots S_{i_k}$, and such that the volume defined by the intersection of the $k+1$ corresponding balls is non empty and is bounded by exactly $k+1$ spherical caps. For each sphere of that tuple, e.g. S_{i_l}, the part of S_{i_l} contained in all other spheres of the tuple defines a spherical cap of order k denoted $O_k(S_{i_l}, \{S_{i_1}\ldots S_{i_{l-1}}, S_{i_{l+1}}\ldots S_{i_k}\})$. A $(k+1)$-tuple then defines $k+1$ spherical caps of order k. The order 0 spherical cap of S_i, $O_0(S_i)$, is S_i exposed surface. The measures used in this study are the areas of the O_k surfaces. Figure 1 shows the 3 sphere case (n=2) with surfaces of order 0, 1 and 2 on S_0. An elaborate strategy to compute all $O_k(S_j, \{S_{j_1}, \ldots, S_{j_k}\})$ consists in retrieving intersection pairs of spheres first, and then computing the surface arrangements for each sphere [6]. The complexity of the arrangement construction is $\mathcal{O}((n+p)\log n)$, n being the number of spheres and p the number of intersections among these spheres. Our robust implementation is based on the 3D Spherical Kernel of CGAL (Computational Geometric Algorithm Library) [7]. The structure of a protein can be described at atomic or coarse

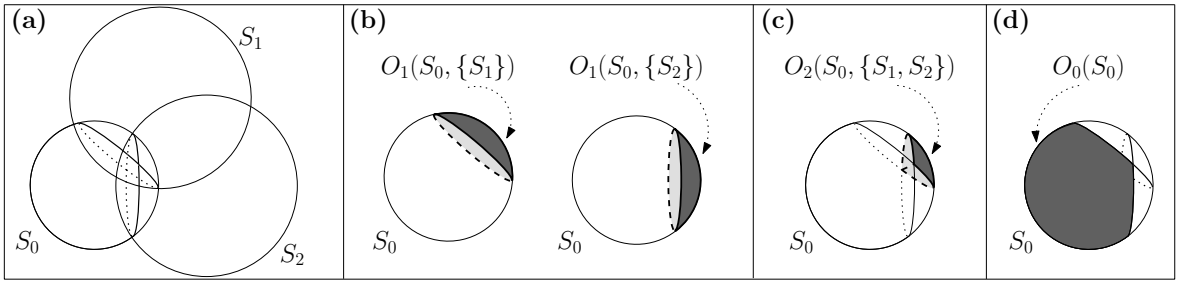


Figure 1. Definition of surface orders. (a) Spheres S_1 and S_2 intersect S_0; (b) Order 1: two surfaces of order 1 are defined by the intersections of both S_1 and S_2 on S_0; (c) Order 2: a surface area of order 2 obtained by the intersection of S_1 and S_2 in S_0; (d) Order 0: the exposed surface of S_0.

grained level using this construction. For an atomic description, O_0 is a description of the exposed surface of the protein, being either the Van der Waals surface when using the Van der Waals atomic radii or the solvent accessible surface when increasing those radii by the radius of a solvent molecule (usually 1.4Å). In this study, we use a coarse-grained representation with one sphere per residue. The center of the sphere is taken as the heavy atom closest to the center of mass of the side chain

including the Cα atom. The radii of the residues were taken from Levitt [14] and increased by 3.5Å in an arbitrary way (trial and error procedure). This value is expected to capture short and mid range interactions. For practical reasons, we limited our study to the areas of the surfaces of order up to 4 (interaction up to 5-sphere). Here the computation of the arrangement and the measure of surface areas takes an average of 10 seconds per protein structure (including all orders).

Datasets. To derive the scoring functions and assess their performance, we need a protein structure set and a set of decoys. High quality of both sets is essential if we are to obtain good quality potentials and scoring functions. For the scoring function construction, we used the dataset from Summa et al.[24] initially designed for refinement procedures. We used a subset of 66 structures corresponding to non-truncated structures, and 729 decoys for each of these structures. To assess the quality of the scoring function construction we performed a 6-fold validation in each function setting using the Summa et al. dataset. We also assessed the scoring function performance on the *lattice_ssfit* dataset from the *decoys'R us* database [25] containing eight proteins, with 2000 decoys for each.

Parameters. For each residue in a structure, the areas of all surface of order 0, 1, 2, 3 and 4 are computed. Both the value of the surface area of the spherical cap and its proportion relative to the total surface area of the sphere are used. To reduce the number of descriptors and see the influence of the residue physical and chemical properties, two types of binning were performed. The first one contains the 20 amino acids binned in 2 groups: (AGCTVILMFWY), (PSNQDERKH) and the second one the amino-acids binned in 6 groups: (FWY), (ILMV), (HKR), (DE), (NQ), (ACGPST).
For each surface of order i, with k amino acid types the number of descriptors is $N_{i,k} = k\binom{i+k-1}{i}$. Considering the large number of descriptors when the order is high, we limited our analysis to order 4, leading to no more than 756 descriptors for the intersection of 5 spheres when using 6 residue types.

Building and Evaluating the Scoring Function. For each descriptor, each value of the surface area (or its relative proportion) is measured. This is the set of observed values obs. Following the RAPDF strategy [22], the ensemble of all the measures can be defined as the reference state ref. A knowledge-based potential function can then be derived by: $E = -kT \log\left(\frac{p_{obs}}{p_{ref}}\right)$. In what follows, we will only consider the potential in its reduced form $\mathcal{S}$, with $\mathcal{S} = -\log\left(\frac{p_{obs}}{p_{ref}}\right)$.
In contrast to the usual knowledge-based potential construction, p_{obs} and p_{ref} are not obtained with a specific binning size. A kernel density estimation is performed on the data using a Gaussian kernel and the Sheather and Jones bandwidth estimation technique [23]. The data are normalized and the log odds is obtained analytically. We built 5 types of potential functions (for each order i), by summing the corresponding descriptors: $\mathcal{S}_i = -\sum_{j=1}^{N_{i,k}} \log\left(\frac{p_{obs_j}}{p_{ref}}\right)$ with $i \in [\![0, 4]\!]$. One questionable approximation is whether the influence of a residue on another has the same weight independently of its relative position. This normalization and the reference state issue have been widely discussed [26]. We kept the simpler model for practical reasons. Also due to the difference of amplitude between the different terms (see section 3), they cannot be simply added to build a combined potential function. This would require a weighting scheme not addressed here.

3 Results and Discussion

Measurements and Potential Construction. Our potential is derived from 66 structures taken from the Summa et al. dataset. When considering groups of 2 residue types, there are between 3000 and 3 million values for each term of the potential. When considering groups of 6 residue types there are between 1000 and 41000 values. Some group types are more common than others, in that hydrophobic residues have a tendency to be buried and interacting less with hydrophobic residues.

Lack of sufficient data prevent us from treating individual residue types and is why we choose a coarse grained model representation (we have 2 or 6 residue types, which is much less than the 167 atom types used in the ENCAD [24] and RAPDF [22] atomic potentials).
For each residue, the equivalent of the knowledge-based potential of mean force, can be defined for surfaces of order 0 through 4. Unlike normal Lennard-Jones interactions or potentials of mean force, our potentials are not repulsive at the origin but at high value. This corresponds to the fact that the residues represented as spheres balls cannot be too close to each other.

Function Evaluation and Native Structure Identification. To quickly evaluate the relative performance of our 5 potential functions, we used the normalized rank of the native structure, i.e. the rank of the native structure divided by the total number of structures (native and decoys) considered: the lower the value, the better the performance. Table 1 summarizes results for decoys from the Summa et al. and the *lattice_ssfit* datasets. For the Summa et al. dataset, 6-fold cross validation was performed for up to order 2 surfaces when using 6 residue groups.
Results show that the exposed surface area (order 0) is not a sufficiently strong score to be able to distinguish the native structure. For order 1 surface, which is related to inter residue distance (and would be totally equivalent if the residue radii were identical), we obtain relatively good performance. Order 2 surface, which corresponds to 3-body interactions, performs slightly less well than order 1 surface but still shows discrimination power. Orders 3 and 4 surfaces, which correspond to multi-body interactions show no discrimination power. Overall, they perform no better than random selection in the Summa et al. dataset, but for some specific structure examples, they appear to give relatively good results (see figure 2).
As may be expected, the 6 residue group functions perform better than the 2 residue group functions, indicating that more than hydrophobic effects are involved in protein structure stabilization. It was not possible to decide whether the surface area or its relative proportion perform best as they seemed to do equally well.
The overall performance is much better for the *lattice_ssfit* dataset. This is to be expected as the Summa et al. decoys were built for refinement purposes and are all near-native structures, basically ranging from 0.02 to 3 Å RMSD. It is thus difficult to select the native structure, especially when we are using a single point for each residue and omitting over 90% of the atoms to make our coarse-grained models. The *lattice_ssfit* dataset contains decoys that have a different fold from the native structure, with RMSD values ranging from 4 to 16 Å.

Summa et al.		$\mathcal{S}_0$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$
2 groups	area	0.438 ± 0.309	0.192 ± 0.226	0.217 ± 0.219	0.450 ± 0.296	0.486 ± 0.303
	proportion	0.427 ± 0.280	0.296 ± 0.264	0.256 ± 0.251	0.464 ± 0.297	0.486 ± 0.305
6 groups	area	0.409 ± 0.298	0.137 ± 0.166	0.205 ± 0.22	0.460 ± 0.295	0.495 ± 0.306
		(0.480 ± 0.304)	(0.190 ± 0.204)	(0.239 ± 0.237)		
	proportion	0.432 ± 0.287	0.095 ± 0.151	0.285 ± 0.279	0.505 ± 0.302	0.524 ± 0.307
		(0.474 ± 0.279)	(0.202 ± 0.229)	(0.307 ± 0.288)		
lattice_ssfit		$\mathcal{S}_0$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$
2 groups	area	0.166 ± 0.206	0.034 ± 0.063	0.042 ± 0.073	0.16 ± 0.200	0.228 ± 0.272
	proportion	0.212 ± 0.248	0.069 ± 0.104	0.040 ± 0.063	0.199 ± 0.230	0.210 ± 0.256
6 groups	area	0.193 ± 0.197	0.023 ± 0.034	0.034 ± 0.053	0.194 ± 0.220	0.227 ± 0.277
	proportion	0.238 ± 0.327	0.002 ± 0.002	0.071 ± 0.132	0.227 ± 0.265	0.242 ± 0.285

Table 1. Normalized ranks of the native structure on the Summa et al. and on the *lattice_ssfit* datasets. Results from the 6-fold cross-validation are in parentheses.

Decoys Evaluation and Comparison With Previous Work. The normalized ranking of the native structure does not give insight on how the potential performs on near native decoys and thus whether it could be used for structure refinement. For all the protein structures evaluated, plots of score vs. RMSD were drawn. Some examples are presented in figure 2.
The exposed (order 0) surface area sometimes indicates the best structure but cannot be used as a selection criterion. In most cases the potentials for order 1 and 2 surface are able to identify and correctly rank near native structures. This is very clear for the Summa et al. dataset: plots show funnel-like shapes (high correlation between score and RMSD), characterizing good structure ranking in a refinement setting. Interestingly for some examples, order 3 and 4 potentials also show good results on the same dataset. Due to the high RMSD values of the *lattice_ssfit* dataset, the structure selection is less clear but still representative.
To compare to a recent study from Bhattacharyay et al. [5], we computed the $Zscore$, the $logPB1$ and the $logPB10$ for the *lattice_ssfit* dataset at order 1 (see [5] for definitions). We obtained average value -1.45, -0.28 and -1.42 respectively. Overall the Bhattacharyay et al. potential performs better than our best potential (-4.06, -0.41, -1.55 respectively). This may be due to the fact that their training dataset is much larger, that they look at smaller spheres or that they normalize surface areas at a single sphere level. These parameters still have to be investigated. We also expect to get better results by combining different terms of surface orders. Our study could then be extended to the whole *decoys'R us* dataset and other decoy datasets.

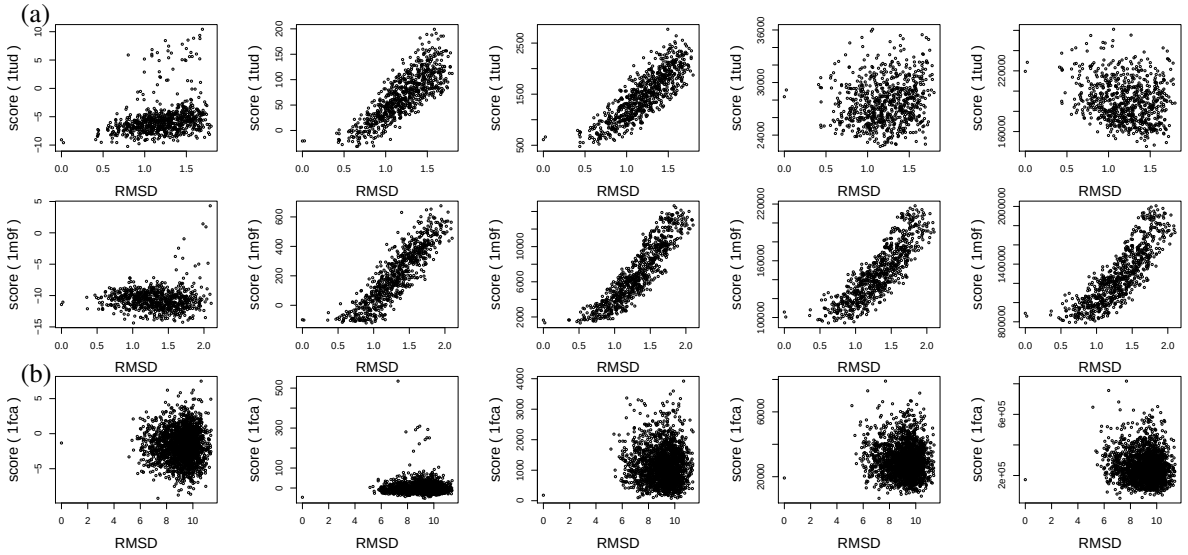


Figure 2. Examples of score vs. RMSD plots. (a) From Summa et al. dataset. (b) From the *lattice_ssfit* dataset.

4 Conclusion and Perspectives

The spherical arrangements leading to surface area measurements have proven to be a good encoding of multi-body contacts. We have shown that it is possible to derive a knowledge based potential that is able to rank correctly native and near native structures in a coarse-grained setting. Improvements over conventional very-well optimized distance-based potentials are small. Further optimization is needed, especially to combine different order scoring functions. This potential is a brand-new type, does not use distance as the primary parameter, is differentiable and could be used for minimization purposes. Combined with an atomic potential for high-resolution refinement, it could also improve structure prediction and model selection.

References

[1] Nataraj Akkiraju and Herbert Edelsbrunner. Triangulating the surface of a molecule. *Discrete Appl. Math.*, 71(1-3):5–22, 1996.

[2] R. L. Baldwin. Making a network of hydrophobic clusters. *Science*, 295(5560):1657–8, 2002.

[3] A. Ben-Naim. Statistical potentials extracted from protein structures: Are these meaningful potentials? *The Journal of Chemical Physics*, 107(9):3698–3706, 1997.

[4] J. Bernauer, J. Aze, J. Janin, and A. Poupon. A new protein-protein docking scoring function based on interface residue properties. *Bioinformatics*, 23(5):555–62, 2007.

[5] A. Bhattacharyay, A. Trovato, and F. Seno. Simple solvation potential for coarse-grained models of proteins. *Proteins*, 67(2):285–92, 2007.

[6] F. Cazals and S. Loriot. Computing the arrangement of circles on a sphere, with applications in structural biology. *Computational Geometry : Theory and Applications*, (in press), 2008.

[7] Pedro M. M. de Castro, Frédéric Cazals, Sébastien Loriot, and Monique Teillaud. Design of the CGAL 3D spherical kernel and application to arrangements of circles on a sphere. *Computational Geometry: Theory and Applications*, (in press), 2008.

[8] M. Delarue and P. Koehl. Atomic environment energies in proteins defined from statistics of accessible and contact surface areas. *J Mol Biol*, 249(3):675–90, 1995.

[9] B. N. Dominy and C. L. Brooks. Identifying native-like protein structures using physics-based potentials. *J Comput Chem*, 23(1):147–60, 2002.

[10] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319(6050):199–203, 1986.

[11] H. H. Gan, A. Tropsha, and T. Schlick. Lattice protein folding with two and four-body statistical potentials. *Proteins*, 43(2):161–74, 2001.

[12] T. Lazaridis and M. Karplus. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol*, 10(2):139–45, 2000.

[13] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, 55(3):379–400, 1971.

[14] M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*, 104(1):59–107, 1976.

[15] H. Lu and J. Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44(3):223–32, 2001.

[16] J. Maupetit, P. Tuffery, and P. Derreumaux. A coarse-grained protein force field for folding and structure prediction. *Proteins*, 69(2):394–408, 2007.

[17] B. J. McConkey, V. Sobolev, and M. Edelman. Discrimination of native protein structures using atom-atom contact scoring. *Proc Natl Acad Sci U S A*, 100(6):3215–20, 2003.

[18] L. A. Mirny and E. I. Shakhnovich. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol*, 264(5):1164–79, 1996.

[19] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, 256(3):623–44, 1996.

[20] B. Park and M. Levitt. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol*, 258(2):367–92, 1996.

[21] J. Qiu and R. Elber. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins*, 61(1):44–55, 2005.

[22] R. Samudrala and J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol*, 275(5):895–916, 1998.

[23] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690, 1991.

[24] C. M. Summa and M. Levitt. Near-native structure refinement using in vacuo energy minimization. *Proc Natl Acad Sci U S A*, 104(9):3177–82, 2007.

[25] Y. Xia, E. S. Huang, M. Levitt, and R. Samudrala. Ab initio construction of protein tertiary structures using a hierarchical approach. *J Mol Biol*, 300(1):171–85, 2000.

[26] H. Zhou and Y. Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*, 11(11):2714–26, 2002.

Présentations courtes

EuGène Maize : a gene prediction web tools for maize

Pierre Montalent[1], Johann Joets[1]

[1] UMR de Génétique Végétale, INRA, Univ Paris-Sud, CNRS, AgroParisTech,
F-91190 Gif-sur-Yvette, France
{montalen, joets}@moulon.inra.fr

Abstract: *The complete sequence of the maize genome is about to be delivered. The prediction of the gene content of these sequences will be the first step to unravel genome organization. EuGène, is an ab initio prediction software, that can in addition combine several sources of evidence. It needs to be trained with well characterized sequence sets. Quality of prediction is dependant of sequence training set quality. We show here that EuGène Maize overperform Twinscan predictor (by 24% for specificity and by 14% for sensitivity of gene prediction) EuGène Maize is available as a web site at http://genome.jouy.inra.fr/eugene/cgi-bin/eugene_form.pl.*

Keywords: Genomics, gene prediction, structural annotation, zea mays.

1 Introduction

L'afflux de séquences génomiques de maïs lié aux nouveaux programmes de séquençage massifs implique la mise au point d'un outil d'annotation utilisable par les groupes à l'origine de ses données (http://www.maizesequence.org/index.html). Le logiciel EuGène [1] a été choisi pour la précision de ses prédictions, inhérente à sa capacité d'intégration d'informations issues de resources multiples (Blast, GenomeThreader, …). Il inclut un prédicteur *ab initio* qui doit être entraîné en utilisant des séquences de gènes dont la structure est finement charcatérisée. Après entraînement et optimisation d'EuGène, la sensibilité et la spécificité des prédictions ont été mesurées. Ce logiciel est accessible à la communauté sous la forme d'un site internet (http://genome.jouy.inra.fr/eugene/cgi-bin/eugene_form.pl).

2 matériels et méthodes

6700 séquences génomiques de maïs et environ 5500 séquences de transcrits étiquetés comme *full length* ou *complete* ont été téléchargées depuis GenBank (http ://www.ncbi.nlm.nih.gov/).

Un pipeline automatique a été développé pour construire le jeu d'entrainement et comprend les étapes suivantes. Les ADNc sont nettoyés par Seqclean [2] couplé à la base Univec [3]. La redondance est éliminée en alignant les ADNc entre eux avec le logiciel Blast. Les couples ADNc/ADNg sont ensuite construits en alignant les ADNc restants sur les séquences génomiques avec le logiciel BLAT [4]. Enfin, la dernière étape est la détermination de la position de la séquence

codante avec GenomeThreader [5] par alignement des protéines de maïs et de riz sur les couples ADNc/ ADNg. Chaque couple est ensuite expertisé manuellement. Certains couples rejetés par le pipeline, ont été réintégrés au jeu d'entrainement final après correction. Le jeu d'entraînement a été divisé en trois lots destinés à l'entraînement, l'optimisation et la validation d'EuGène selon la documentation du logiciel.

3 Résultats

Le jeu de séquence d'entraînement expertisé comprend 352 couples ADNc/ADNg dont 251 sont finalement utilisés après exclusion des séquences comportant des sites d'épissage non canoniques. L'ensemble des prédicteurs intégrés dans EuGène Maize sont GenomeThreader (alignements épissés), BlastX, SpliceMachine [6] (predicteur des sites de démarrage de traduction et des sites donneurs et accepteur d'épissage), Fgenesh [7] (prédiction de gènes *ab initio*). Les banques utilisées avec les logiciels d'alignement de séquences comprennent selon les cas : les ARNm « full-length » de maïs (GenBank), les ARNm de riz (RAPdb), les contigues d'EST de maïs (PUT [8]). les protéines de maïs (Swiss-Prot), de riz (RAPdb), d'*Arabidopsis* (TAIR6).

La qualité de prédiction est estimée par la mesure de la spécificité et de la sensibilité des prédictions. Ces mesures (table 1) montrent qu'EuGène Maize offre de meilleurs résultats que les autres prédicteurs publiés entaînés pour le maïs ou les monocotyledones [9].

	Exon spé.	Exon sens.	Gène spé.	Géne sens.
Fgenesh	50.6	72.6	24.7	33.7
Twinscan 3.5	57.8	83.9	33	51
Eugene 3.4	88	91.1	56.8	65.1

Table 1. Qualité de prédiction pour le maïs (spécificité, sensibilité).

EuGène Maize peut être utilisé en ligne. Les resultats sont retournés par courrier électronique sous formes de fichiers qui peuvent être visualisés avec un navigateur WWW.

Remerciements

Ce travail a été financé par le programme ANR/génoplante BIEP et a bénéficié du support technique des plates-formes bioinformatiques MIGALE (C. Caron, Jouy en Josas) et URGI (D. Steinbach, Versailles) ainsi que des conseils de Jêrome Gouzy et Thomas Schieix (INRA Toulouse).

References

[1] Foissac S, Schiex T, Integrating alternative splicing detection into gene prediction. BMC Bioinformatics : 6:25 - Feb 10 2005
[2] SeqClean. http://www.tigr.org/tdb/tgi/software/
[3] Univec http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html
[4] Kent WJ. BLAT—the Blast-like alignment tool. Genome Res. 12:656–664. 2002
[5] G. Gremme, V. Brendel, M. E. Sparks and S. Kurtz, Engineering a software tool for gene structure prediction in higher organisms. Information and Software Technology 47 965–978. 2005
[6] S. Degroeve, Y. Saeys, B. De Baets, P. Rouze and Y. Van de Peer Bioinformatics 21 :1332-8. 2005.
[7] A.A. Salamov and V. V. Solovyev, Ab initio gene finding in Drosophila genomic DNA. Genome Res. 10:516–522. 2000
[8] PUT. http://www.plantgdb.org/prj/ESTCluster/PUT_procedure.php
[9] Barbazuk, Brad Yu, Yan Zhang, Chenhong Brent, Michael R, Maize trained TWINSCAN and ab initio gene finding in maize, 50th Annual Maize Genetics Conference, February 27 – March 1, 2008 Washington, D.C.

Intégration automatique d'une ontologie de domaine dans un annuaire BioMoby

Julien Wollbrett, Pierre Larmande et Manuel Ruiz

CIRAD-BIOS, UMR Développement et Amélioration des Plantes
TA A-96/03, Avenue Agropolis 34398 Montpellier Cedex 5
[julien.wollbrett,pierre.larmande,manuel.ruiz]@cirad.fr

Abstract: *Crop data management systems deal with data integration problems. The interoperability of these systems could be increased by sharing plant ontologies through BioMoby web services. However, speed up datatypes and services registration is lacking. We developed a plug-in for Protégé named BioMoby Converter in order to register an OWL ontology into a BioMoby registry.*

Keywords: BioMoby, semantic web, ontologies, data integration.

1. Introduction

Dans le cadre de projets internationaux, des quantités souvent considérables de données sont générées présentant par nature une large hétérogénéité et une forte évolutivité. Afin de valider des hypothèses un biologiste peut avoir besoin d'un accès unique et transparent à des sources multiples, réparties et hétérogènes, essentiellement accessibles par le Web. Les propositions actuelles, les plus à même de répondre à ce défi, relèvent de la thématique d'intégration et médiation de données. Comme l'ont décrit des articles de synthèse un grand nombre de solutions et de systèmes existent en bioinformatique [1,2]. Notre travail s'inscrit dans ce contexte, et plus particulièrement dans le cadre d'un projet international nommé Generation Challenge Programme (GCP). Le projet GCP a pour objectif d'intégrer des données génétiques de plantes d'intérêts agronomiques à l'aide d'une architecture de médiation appelé GCP Pantheon [3]. Certains adaptateurs, nécessaires pour faire le pont entre les sources de données et GCP Pantheon, utilisent les Services Web BioMoby [4]. Ces Services Web s'appuient sur l'ontologie de domaine GCP afin de communiquer avec la plateforme. Actuellement l'étape d'enregistrement de Services Web BioMoby compatibles avec GCP Pantheon n'est pas automatisée. Elle requiert soit l'utilisation d'API BioMoby (Perl ou Java), qui ne reste accessible qu'aux programmeurs, soit l'utilisation d'une interface graphique BioMoby Dashboard, qui est limitée à des actions unitaires. Afin de rendre cette étape automatique, un plugin nommé BioMoby Converter a été réalisé sous Protégé [5]. BioMoby Converter a pour objectif de donner à tout utilisateur la possibilité d'enregistrer automatiquement une ontologie de domaine vers un annuaire BioMoby. Cette ontologie sera ainsi utilisable pour la création de Services Web sémantiques sous BioMoby. Dans le cas de notre travail, nous nous appuyons sur l'ontologie de domaine du GCP.

2. Méthodologie

Deux formats principaux sont généralement utilisés pour la création d'ontologies, à savoir (i) le format OBO (Open Bioinformatical Ontology) [6] développé par la communauté Gene Ontology,

qui est un standard en bioinformatique, (ii) le format OWL [7] qui est un standard du web sémantique.

Protégé est un logiciel de création d'ontologies OWL qui couplé avec le plugin OboConverter [8], permet de passer du format OBO au format OWL. En étendant les fonctionnalités de Protégé par l'implémentation d'un nouveau plugin, nous rajoutons la possibilité d'enregistrer automatiquement une ontologie ou une partie d'ontologie OBO ou OWL sur un annuaire BioMoby. Nous rajoutons également la possibilité d'exporter une ontologie d'un annuaire BioMoby vers un fichier OWL.

3. Résultats

L'intégration ou la mise à jour d'une ontologie dans un annuaire BioMoby est largement facilitée. L'utilisateur peut éditer une ontologie avec Protégé, puis BioMoby Converter permet de sélectionner l'annuaire BioMoby dans lequel il souhaite enregistrer son ontologie et de valider son choix. BioMoby Converter se charge ensuite automatiquement de récupérer les informations nécessaires dans le fichier OWL, de trier les données et de les enregistrer sur l'annuaire BioMoby préalablement sélectionné. L'utilisation de BioMoby Converter permet un gain de temps conséquent.

4. Limites et perspectives

Les limitations de BioMoby Converter sont dues à une perte de sémantique lors de la transformation d'une ontologie au format OWL vers les ontologies de BioMoby. En effet, dans le cas d'une ontologie BioMoby, les relations possibles sont ISA (relation de parenté), HAS (relation d'agrégation) et HASA (relation d'agrégation de cardinalité 1). Il n'est donc pas envisageable de vouloir retranscrire l'intégralité de la sémantique présente dans une ontologie OWL en utilisant les 3 relations disponibles dans BioMoby. BioMoby montrant ses limites sémantiques, nous envisageons d'étendre l'utilisation du plugin au projet SSWAP [9].

References

[1] T. Hernandez et S. Kambhampati, "Integration of Biological Sources: Current Systems and Challenges Ahead.," *SIGMOD Record*, vol. 33, 2004, pp. 51-60.

[2] L.D. Stein, "Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges," *Nat Rev Genet*, vol. 9, 2008, pp. 678-688.

[3] R. Bruskiewich, M. Senger, G. Davenport, M. Ruiz, M. Rouard, et al., "The Generation Challenge Programme Platform: Semantic Standards and Workbench for Crop Science," *International Journal of Plant Genomics*, 2008.

[4] M.D. Wilkinson et M. Links, "BioMOBY: an open source biological web services proposal.," *Brief Bioinform*, vol. 3, Déc. 2002, p. 331—341.

[5] N.F. Noy, R.W. Fergerson, et M.A. Musen, "The knowledge model of Protege-2000: Combining interoperability and flexibility.," *2th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000) Juan-les-Pins, France,*, 2000.

[6] B. Smith et al., "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat Biotech*, vol. 25, Nov. 2007, pp. 1251-1255.

[7] M. Dean et G. Schreiber, *OWL Web Ontology Language Reference*, 2004.

[8] D.A. Moreira et M.A. Musen, "OBO to OWL: a protege OWL tab to read/save OBO ontologies," *Bioinformatics (Oxford, England)*, vol. 23, Juillet. 2007, pp. 1868-70.

[9] D. Gessler, "SSWAP: Simple Semantic Web Architecture and Protocol."

Estimation of sequence errors and capacity of genomic annotation in transcriptomic and DNA-protein interaction assays based on next generation sequencers

Nicolas Philippe[1,2], Anthony Boureux[2], Laurent Bréhélin[1], Jorma Tarhio[3], Thérèse Commes[2] and Eric Rivals[1]

[1] Laboratoire d'Informatique, de Robotique et de Microélectronique,
Université de Montpellier II, UMR 5506 CNRS ; 161 rue Ada, 34392 Montpellier 05, France.
{nphilippe, rivals}@lirmm.fr
[2] Groupe d'études des transcriptomes, Université de Montpellier II,
Place Eugène Bataillon, 34095 Montpellier 05, France.
{Anthony.Boureux, commes}@univ-montp2.fr
[3] Helsinki University of Technology, P.O. Box 5400,
FI-02015 HUT, Finland.

Abstract: *Ultra-high throughput sequencing allow to analyse on a genome-wide scale the transcriptome or the interactome at unprecedented depth. These techniques yield short sequence reads that are then mapped on a genome sequence to predict putatively transcribed or protein-interacting regions. We argue that factors such as false locations, sequence errors, and read length impact on the mapping prediction capacity of these short reads. Here we suggest a computational approach to measure those factors and analyse their influence on both transcriptomic and epigenomic assays. This investigation provides new clues on both methodological and biological issues. Following our procedure, we obtain less than* 1% *of false positives among genomic locations. Therefore, even rare signatures, if they are mapped on the genome, should identify biologically relevant regions. This indicates that digital transcriptomics may help to characterise the wealth of yet undiscovered, low abundance transcripts.*

Keywords: Transcriptomics, High-throughput Sequencing, Genomics, Annotations, Perfect matching.

Next-generation sequencing technologies, able to yield millions of sequences in a single run, allow to interrogate the transcriptome or to assay protein-DNA interactions (by Chromatin Immuno-Precipitation and sequencing (also called ChIP-seq)) at a genome-wide scale. These assays yield short sequences ($<$40 bp), also called *tags*, that need to be mapped to the genome sequence. To each tag is associated the number of times the same sequence has been experimentally detected: its *occurrence number*. For transcriptomic assays, for instance, a tag with a high occurrence number likely is the biologically valid signature of an abundant transcript, while a tag with a low occurrence number may either result from a sequencing error or identify a rare RNA.

The mapping is a compulsory step to first predict, and then annotate regions of interest on the genome. Usually, only genomic locations that are unambiguously mapped by a tag are further analysed. Those high-throughput assays are intended to predict a maximum number of genomic locations of interest. Obviously, this induces a balance between the number of mapped tags and the number of tags that map a unique genomic location, and this balance is controlled by the tag length. The

sequencing technique generally dictates the tag length. Nevertheless, once a certain length is sequenced (e.g., 36 bp with a Solexa/Illumina 1G machine) it is still possible to map only sub-parts (a prefix, a suffix, a substring) of the tags to the genome, thereby artificially reducing the tag length and modifying the balance.

Presently, we lack a statistical method to evaluate the influence of the tag length on the capacity of prediction for different assays and sequencing techniques, as well as the importance of sequence errors. Our contribution is threefold. Based on word statistics, we design a program that computes the theoretical probability of mapping a genomic location by chance for a given tag length.

For this, we computed in function of tag length the probabilities of a tag to be mapped on the genome at least once, and to match a unique location under a Bernoulli model. We approximate the law of these probabilities using the guaranteed Poisson approximation. Using an efficient algorithm to map short tags on complete genome sequence, we investigate how the prediction capacity varies with tag length. Finally, we propose a method to estimate the probability of a tag to be altered by a sequencing error. We apply it to derive a probability of having an erroneous nucleotide at a given position in the tag for the Sanger and Solexa sequencing techniques, and for both transcriptomic and ChIP-seq experiments. We investigate on real data sets how the number of uniquely predicted genomic regions varies with tag length and background distribution. This enables a technical assessment of such assays and the indirect measurement of the impact of some biological phenomena (e.g., the number of reads affected by SNPs). By applying this procedure, we were able to estimate:

1. that 4.6% of reads are affected by SNPs.
2. the nucleotide error probability is low, and it significantly increases with the position in the sequence.
3. by choosing a read length above 19 bp, we practically eliminates the risk of finding irrelevant positions.
4. the number of uniquely mapped reads decreases with sequences above 20 bp.
5. we obtain 0.6% of false positives among genomic locations.

Our analysis delivers the first estimates of sequence error rate for transcriptomic and DNA-protein interaction assays based high-throughput sequencing.

Acknowledgements

The authors wish to thank Skuld-Tech for the SAGE-Solexa library. This work was supported by "La ligue régionale contre le Cancer" Languedoc Roussillon and the BioMIPS grants of the "Université de Montpellier 2". We acknowledge the support of the Languedoc-Roussillon Plateform MontpellierGenomix and the ATGC plateform at LIRMM.

Oenococcus oeni genome plasticity associated with adaptation to wine, an extreme ecological niche

Elisabeth Bon[1], Arnaud Delaherche[2], Eric Bilhère[2], Cécile Miot-Sertier[2], Pascal Durrens[1]
Antoine de Daruvar[1], Aline Lonvaud-Funel[2] and Claire Le Marrec[2]

[1] LaBRI-Université Bordeaux 2 (CNRS-Université de Bordeaux),
351, cours de la Libération 33405 Talence cedex France
elisabeth.bon@labri.fr
[2] UMR INRA 1219, ISVV (Université de Bordeaux)
210, Chemin de Leysotte, 33882 Villenave d'Ornon Cedex France

Abstract: *Genomic subtractive hybridization between two Oenococcus oeni isolates with diametrically opposite oenological aptitudes was used to elucidate part of the genetic bases of this intraspecies diversity and to identify novel genes involved in adaptation to wine.*

Keywords: Comparative genomics, subtractive hybridization, accessory genome, horizontal gene transfer, *Oenococcus oeni,* tolerance to wine.

1 Introduction

Oenococcus oeni, a lactic acid bacteria, is part of the natural microflora of wine and related environments, and is the main agent of the malolactic fermentation (MLF). *O. oeni* strains are well-known for their considerable natural phenotypic variations in terms of tolerance to harsh wine conditions and malolactic activity. These variations are thought to account for the unpredictability of MLF.

2 Strategy

In vitro subtractive hybridizations (SH) were performed between two *O. oeni* isolates with opposite oenological potential (OP), the IOEB-SARCO-1491 and the IOEB-8413 strains having respectively, a high potential (HP) and a low potential (LP). The obtained SH fragments were tested for their strain-specificity, annotated and ordered to reconstitute the original genomes, through an in-house bioinformatics strategy [1]. The association of selected SH sequences with adaptation to wine was further assessed by PCR-screening a collection of *O. oeni* strains with characterized OP, and by studying gene expression patterns following exposure to several common stresses in wine.

3 Results and Discussion

SH revealed 126 specific open reading frames (ORFs) in the HP strain identified as divergent or novel sequences. They were found to represent ~2% of the total number of ORFs present in the complete *O. oeni* sequenced genomes [2,3], and were added to the global repertoire of *O. oeni.* A large proportion was shown to resemble genes involved into carbohydrate transport and metabolism, cell wall/membrane/envelope biogenesis, replication, recombination and repair.
Six major regions (I to VI) of genomic plasticity were identified. Their analysis suggested that limited recombination (region III) as well as punctual or regional insertion-deletion (region II, 30 ORFs) events play a role in creating diversity in *O. oeni*. Mobile genetic elements (MGEs) and "alien" genes originating from other bacteria were found to significantly contribute to the HP strain genomic specificity. Among these "alien" sequences, 21 genes were physically clustered in four loci (I, IV, V, VI) and displayed a restricted distribution among our strain panel. These data may denote gene acquisition through horizontal gene transfer (HGT). Accordingly, region IV had additional hallmarks of HGT (anomalies in GC content, presence of flanking IS elements).
The analysis of the prevalence of 28 selected sequences representing the six regions of plasticity as well as randomly-selected singletons in the collection of *O. oeni* strains showed a statistically significant positive association between HP strains and the presence of 8 gene sequences residing on regions II, IV, and V. The modification of their expression pattern under exposure to common stresses in wines, clearly confirmed that these genes were of oenological interest.
Five out the eight target genes were clusterized in region IV (~8kb), a genomic islet proposed to originate from HGT, suggesting that the strategy of acquiring genes from other bacteria enhances the fitness of *O. oeni* strains. The structure of this 8-kb adaptative islet was examined in other *O. oeni* isolates. A few strains harbored the islet as part of a larger 24-kb mobile genomic island, also sandwiched between two IS copies. Unexpectedly, in vitro experiments demonstrated that region IV was able to excise in its whole from the chromosome to form fleeting free circular intermediates.

4 Conclusion

This study helped us to identify novel genes having a great oenological interest that could be used as markers to genotype the most performing *O. oeni* strains, which is commercially essential. This project gave us the first clues of the genetic origin of *O. oeni* strain phenotypic diversity. It contributed to partially disclose the *O. oeni* accessory genome repertoire content, consisting in strain-specific loops and pulses, and to disclose the complex origin of genome plasticity that presumably result from a fine equilibrium between clonal divergence and horizontal gene transfert. Finally, this study also helped us to enrich our knowledge on the *O. oeni* pangenome architecture, which is evolutionary essential.

References

[1] E. Bon, A. Delaherche, E. Bilhère, A. de Daruvar, A. Lonvaud-Funel and C. Le Marrec, *Oenococcus oeni* genome plasticity is associated with fitness. Appl. Environ. Microbiol., 75(7):2079-90, 2009.
[2] E. Bon, C. Grandvalet, […], A. Lonvaud-Funel and J. Guzzo, Insight into genome plasticity of the wine-making bacterium *Oenococcus oeni* strain ATCC BAA-1163 by decryption of its whole genome. Proceedings of the 9th Symposium on Lactic Acid Bacteria, FEMS, The Netherlands, ref. A047, 2008.
[3] DA. Mills , H. Rawsthorne, C. Parker, D. Tamir and K. Makarova, Genomic analysis of *Oenococcus oeni* PSU-1 and its relevance to winemaking. FEMS Microbiol Rev., 29(3):465-75, 2005.

Databases of homologous gene families for comparative genomics

Simon Penel[1], Anne-Muriel Arigon[2], Vincent Daubin[1], Pascal Calvat[3], Stéphane Delmotte[1], Jean-François Dufayard[2], Manolo Gouy[1], Guy Perrière[1], Anne-Sophie Sertier[1], Laurent Duret[1]

[1]Laboratoire de Biométrie et Biologie Evolutive, CNRS UMR 5558, UCBL, Lyon I
43, Bd du 11 novembre 1918, 69622 Villeurbanne cedex –France
{penel,daubin,delmotte,mgouy,perriere,sertier,duret}@biomserv.univ-lyon1.fr
[2] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
161 rue Ada, 34392 Montpellier, France
{Anne-muriel.Arigon,Jean-francois.Dufayard}@lirmm.fr
[3] Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules
27 Bd du 11 novembre 1918, 69622 Villeurbanne cedex –France
calvat@cc.in2p3.fr

Abstract: *Comparative genomics is a central step in many sequence analysis studies, from gene annotation and the identification of new functional regions in genomes, to the study of evolutionary processes at the molecular level (speciation, single gene or whole genome duplications, etc.) and phylogenetics. In that context, databases providing users high quality homologous families and sequence alignments as well as phylogenetic trees based on state of the art algorithms are becoming indispensable.*

Keywords: Phylogenomics, databases, lateral gene transfer, molecular evolution.

1 Introduction

Comparative genomics is a central step in many sequence analysis studies, from gene annotationand the identification of new functional regions in genomes, to the study of evolutionary processes at the molecular level (speciation, single gene or whole genome duplications, etc.) and phylogenetics. In that context, databases providing users high quality homologous families and sequence alignments as well as phylogenetic trees based on state of the art algorithms are becoming indispensable. The interest of these databases compared to other databases as COG relies on its phylogenomic approach wich allows to use tree toplogies to retrieve orthologous/paralogous sets of genes.

2 Methods

We developed an automated procedure allowing massive all-against-all similarity searches, gene clustering, multiple alignments computation, and phylogenetic trees construction and reconciliation.

The application of this procedure to a very large set of sequences is possible through parallel computing on a large computer cluster.

3 Results

Three databases were developed using this procedure: HOVERGEN, HOGENOM and HOMOLENS.These databases share the same architecture but differ in their content. HOVERGEN contains sequences from vertebrates, HOGENOM is mainly devoted to completely sequenced microbial organisms, and HOMOLENS is devoted to metazoan genomes from Ensembl. One can use these databases according to the research interest. For example HOGENOM can be used to study gene transfers in bacteria, HOVERGEN to study vertebrates genes and HOMOLENS to study the evolution of animals. Access to the databases is provided through Web query forms, a general retrieval system and a client-server graphical interface. The later can be used to perform tree-pattern based searches allowing, among other uses, to retrieve sets of orthologous genes. The three databases, as well as the software required to build and query them, can be used or downloaded from the PBIL (Pole Bioinformatique Lyonnais) site at http://pbil.univ-lyon1.fr/databases/hogenom.html (and databases/homolens.php, databases/hovergen.php respectively).

4 New ressources for the building of homologous gene family databases

Since the explosion of data generated by sequencing projects of more and more organisms, the maintenance of an increasing number of homologous gene family databases implies to develop new approaches. In the view, we developed a complete genome interactive tank (COGIT), a non-redundant sequence database (BGENR) and its associated BLAST hits database.

Acknowledgements

Calculations have been done at the IN2P3 Computing Center.

ace.map – a comprehensive tool for advanced microarray analysis

Guillaume Brysbaert[1,2], Brice Targat[1,2], Nicolas Tchitchek[1,2], Jose Felipe Golib Dzib[1,2], Christophe Bécavin[1,2], Sebastian Noth[1,2] and Arndt Benecke[1,2]

[1] Institut de Recherche Interdisciplinaire, CNRS USR3078, 50 avenue de Halley, BP 70478, 59658 Villeneuve d'Ascq Cedex France
guillaume.brysbaert@iri.univ-lille1.fr
[2] Institut des Hautes Etudes Scientifiques, CNRS, 35 route de Chartres, 91440 Bures-sur-Yvette, France

Abstract: *We describe here a novel stand-alone JAVA tool for the systematic and advanced statistical analysis of high-density transcriptome microarray data: ace.map. It permits not only to lead a complete analysis through standard statistics tools like other applications (normalization, weighted-merging, subtraction-profiling, clustering…), but also to use new analysis tools like kinetic analyses, multidimensional scaling of biological conditions and ontology-related analyses. We have proven efficiency of our application in several research projects on AIDS, Epstein-Barr virus, and breast cancer, and expect its ergonomy and completeness will stimulate the scientific community in taking advantage of this resource to mine their transcriptome data.*

Keywords: microarray data analysis, transcriptome, clustering, single value decomposition, multidimensional scaling, kohonen maps, classification, ontology.

1 Introduction

Considering the ever increased number of high-density microarray platforms that has emerged during the last decade, an important need for complete, standardized, and efficient data analysis has evolved. Coupled with the observation that the AB1700 platform we use produces double-lognormal data signal distributions [1] which is in stark contrast to other technologies, we decided to develop *ace.map*, a new efficient Java tool for microarray analysis. Our software is built to be able to deal with both single and double lognormal distributed data, permits adapted classic statistical analyses on microarray data, and includes new methodologies developed in our group.

2 ace.map application

This application leans on a strong structure for data: the annotation of each experiment is enclosed to its data and both are gathered in an archive. Annotation information follows MIAME recommandations [2]. In the same way, for each analysis done, an annotation is enclosed to the data results, with all the parameters and information associated to the analysis. In that manner, data are easier to understand, easier to mine, easier to share. Compared to other microarray analysis softwares, *ace.map* has been built to include specific tools for double lognormal distributions, and

puts more emphazise on data transparency and sharing.

ace.map allows to treat data with standard statistics tools like weighted merging of technical replicates, but we also adapted some of them. In subtraction profiles of biological replicates, for instance, the NeONORM inter-assay normalization [3] is implemented. Tools for gene signature definition are included. Principal component analysis or single linkage hierarchical clustering are standard applications for data mining. Considering the different data models we developed [1] we also provide data quality analysis based on the intrinsic statistics of the distributions [4].

Our application goes further than standard analysis software, implementing new algorithms to better analyze data in time, to better visualize data and better link them to global regulation networks in cell. First, kinetic analysis based on a Kohonen-map classifier clusters data in function of their topography of expression in time. Second, GEO is an algorithm that allows the use of multidimensional or classical scaling for representation of gene or biological conditions in different space embeddings (typically N-1, 3D, 2D). Finally, LEO is a module that permits advanced ontological analysis using priors on gene expression and co-expression probabilities in order to highlight the pathways, biological processes or molecular functions that are affected in a particular biological condition. It furthermore differs from e.g the Panther tool (http://www.pantherdb.org/) by also considering weights for probes in function of their appartenance to one or several ontologies.

All the data generated via the different analysis modules in the context of a project can be gathered in one global project archive via the application. In that way, complete analyses can be easily shared, rebuilt, and exploited by third parties. This particular data structure has already proven very helpful in the context of different projects using data from different technology platforms [5,6,7,8,9].

3 Conclusion

ace.map is a new integrated tool for advanced microarray analysis that permits to easily mine microarray data and to identify regulation networks. Hypotheses building and testing procedures are provided. We expect that such a complete tool with its simple handling has a place next to existing solutions and its use should contribute to a more integrated and systematic analysis of increasingly heterogeneous and numerous experimental data.

References

[1] Noth, S., et al. (2006) High-sensitivity transcriptome data structure and implications for analysis and biologic interpretation. Genomics, Proteomics & Bioinformatics (GPB) 4:212-229.
[2] Brazma A, et al. (2001) MIAME – toward standards for microarray data. Nat Gen. 2001 Dec;29(4):365-71
[3] Noth, S., et al. (2006) NeONORM provides robustness against asymmetries in comparative transcriptome profiles and avoids false calls. Genomics Proteomics Bioinformatics 4:90-109.
[4] Pellay FX, Brysbaert G, et al. (2009) Quality assessment of transcriptome data using intrinsic statistical properties. (submitted)
[5] Wilhelm E, et al. (2008) TAF6delta controls apoptosis and gene expression in the absence of p53. PLoS One. 3(7): e2721
[6] Wilhelm E, et al. (2008) Determining the impact of alternative splicing events on transcriptome dynamics. BMC Res. Notes. 1:94.
[7] Firlej V, et al. (2008) Reduced tumorigenesis in mouse mammary cancer cells following inhibition of either of the distinct Pea3 and Erm transcription programs. J Cell Sci. 121:3393-402.
[8] Jacquelin B, et al. (2007) Long oligonucleotide microarrays for African Green Monkey gene expression profile analysis. FASEB J. 21:3262-71.
[9] Eilebrecht S, Pellay FX, et al. (2008) EBER2 RNA-induced transcriptome changes identify cellular processes likely targeted during Epstein Barr Virus infection. BMC Res Notes. 2008 Oct 28;1:100.

Crossing genome and transcriptome: deciphering links between structure and function in *Arabidopsis thaliana* genes

Véronique Brunaud[1], Virginie Bernard[1], David Armisén[1], Jean-Philippe Tamby[1], Séverine Gagnot[1], Sandra Derozier[1], Franck Samson[1], Cécile Guichard[1], Marie-Laure Martin-Magniette[1,2], Alain Lecharny[1] and Sébastien Aubourg[1]

[1] Unité de Recherche en Génomique Végétale (URGV) - UMR INRA 1165-CNRS 8114-UEVE, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France.

[2] Unité de Mathématiques et Informatique Appliquées (MIA) - UMR 518 AgroParisTech-INRA, 16 Rue Claude Bernard, 75231 Paris Cedex, France.

(brunaud,bernard,armisen,tamby,gagnot,derozier,samson,guichard,martin, lecharny,aubourg)@evry.inra.fr

Abstract: *One of the challenges of bioinformatics today is to cross and analyse data differing in type, origin and quality in order to increase our knowledge on genomes. We have developed an information system focused on the model plant Arabidopsis and allowing the improvement of the functional annotation of genes by combining transcriptome and structural data.*

Keywords: transcription, annotation, evolution, regulation, plant, database, integration

1 Introduction

Since the sequencing of the *Arabidopsis thaliana* whole genome 8 years ago, gene annotation has been improved through several releases taking advantage of resources as like as expert curation, new genome availability, transcript sequencing projects and gene prediction software improvement. In parallel, several transcriptomic platforms based on DNA chips have been developed and they now give access to several thousands of transcriptomes [1,2]. Nevertheless, only 14% of the Arabidopsis genes have a biological function that was experimentally assessed while around 20% of genes remain without any functional information. In the context of functional genomic projects for plants, we have developed two databases for the management of genomic (FLAGdb^{++}, [3]) and transcriptomic data (CATdb, [4]). This information system allows us to integrate, through holistic approaches, gene models and expression profiles in order to improve the genome annotation and to decipher relationships between the organization, evolution and function of Arabidopsis genes.

2 Results

A combined approach of genome annotation and transcript analysis was firstly performed to identify new genes in the Arabidopsis genome. Probes on the CATMA microarrays were based on the gene models predicted by the EuGène software [5] and 677 probes were located within regions that were considered as intergenic by the official TAIR annotation. The statistical analysis of the results for more than 500 hybridized samples distributed among 12 organs provided an experimental validation for 465 novel genes [6]. These novel genes were characterized by their small size (encoding proteins with an average size of 137 aa) and very specific expression patterns.

Another illustration of the advantage to combine genomic and transcriptomic data is the characterization of a particular class of genes in plants: the unique genes [7]. Despite the major role of duplications in genome evolution, all characterized genomes include unique (single-copy) genes, i.e. genes without apparent paralog. Mining the FLAGdb^{++} database, we identified the unique genes within both Arabidopsis and rice genomes and classified them according to the number of homologs in the alternative species. Unique gene sets share structural features. In particular, the conserved unique gene pairs are characterized by a relatively small protein size, a high intron density, a rare

occurrence of TATA-box and a high occurrence of TELO-box. These structural features predict these genes as preferentially house-keeping genes with a slow evolution. Even if no shared transcription factor binding site (TFBS) can be detected in their promoter, the orthology relationship in Arabidopsis-rice gene pairs was strongly supported by a high conservation of their transcription levels. Furthermore, many unique genes have been conserved in single-copy throughout evolution from Prasinophytes to angiosperms, indicating that the uniqueness is under a strong selective pressure. A high proportion of conserved unique genes was also observed in other life phylums and we showed a link between protein targeting towards plastids and homology with bacterial proteins [7].

The expression of mature transcripts is controlled by the intron-exon structures and by the TFBS content of promoter regions. Also we have initiated a genomic study of the links between the core promoter architecture and gene function in Arabidopsis. Firstly, we identified different motifs with topological features that are strongly similar to canonical TATA-box features suggesting that they are functional motifs. Based on theses sequences, we established a novel classification of promoters and described links between promoter gene classes and the Gene Ontology categories. Secondly, we focused on the house-keeping genes, i.e. the genes expressed in almost all the conditions and organs, and found that TATA-box is under-represented in these genes. This atypical class of genes is also characterized by a compact structure (shorter introns and coding sequences).

3 Conclusion

The CATMA transcriptomes available in the CATdb database are fully independent of other public transcriptome resources. For instance, more than 4000 gene probes (including miRNA genes) are only present on CATMA chips. Using CATMA, it is therefore possible (i) to cross-validate results inferred from other resources [8] and (ii) to improve our Arabidopsis gene knowledge through the ‘guilt by association’ strategy.

References

[1] Graham NS, Broadley MR, Hammond JP, White PJ, May ST. Optimising the analysis of transcript data using high density oligonucleotide arrays and genomic DNA-based probe selection. BMC Genomics. 8:344, 2007.

[2] Allemeersch J, Durinck S, Vanderhaeghen R, Alard P, Maes R, Seeuws K, Bogaert T, Coddens K, Deschouwer K, Van Hummelen P, Vuylsteke M, Moreau Y, Kwekkeboom J, Wijfjes AH, May S, Beynon J, Hilson P, Kuiper MT. Benchmarking the CATMA microarray. A novel tool for Arabidopsis transcriptome analysis. Plant Physiol. 137:588-601, 2005.

[3] Samson F, Brunaud V, Duchêne S, De Oliveira Y, Caboche M, Lecharny A, Aubourg S. FLAGdb++: a database for the functional analysis of the Arabidopsis genome. Nucleic Acids Res. (Database issue) 32:D347-350, 2004. **http://urgv.evry.inra.fr/FLAGdb**

[4] Gagnot S, Tamby JP, Martin-Magniette ML, Bitton F, Taconnat L, Balzergue S, Aubourg S, Renou JP, Lecharny A, Brunaud V. CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. Nucleic Acids Res. (Database issue) 36:D986-990, 2008. **http://urgv.evry.inra.fr/CATdb**

[5] Schiex T, Moisan A, Rouzé P. Eugene, an eukaryotic gene finder that combines several sources of evidence. Lect. Notes Computational Sciences 2066:111-125, 2001.

[6] Aubourg S, Martin-Magniette ML, Brunaud V, Taconnat L, Bitton F, Balzergue S, Jullien PE, Ingouff M, Thareau V, Schiex T, Lecharny A, Renou JP. Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. BMC Genomics. 8:401, 2008.

[7] Armisén D, Lecharny A, Aubourg S. Unique genes in plants: specificities and conserved features throughout evolution. BMC Evol. Biol. 8:280, 2008.

[8] Hughes TR. 'Validation' in genome -scale research. Journal of Biology, 8:3, 2009.

Generalized Peptide Mass Fingerprinting on Whole-Cell HPLC-MS Proteomics Experiments

Pascal Bochet[1], Frank Rügheimer[1], Tina Guina[2], David Goodlett[2], Peter Clote[3], Benno Schwikowski[1].

[1] Laboratoire de Biologie systémique
Institut Pasteur, CNRS URA 2171
25, rue du Docteur Roux 75015 Paris France
pascal.bochet@pasteur.fr
[2] Dpt. of Pediatrics and
[3] Dpt. of Medicinal Chemistry, U. of Washington, Seattle,WA
[4] Biology Dpt., Boston College, Boston MA

Abstract: *We present and evaluate a novel (and, to our knowledge, first) method to identify proteins from large-scale LC/MS proteomics experiments without additional fragmentation information.*

Keywords: Proteomics, Protein identification, Mass Spectrometry, Genomics, HPLC, Retention time.

1 Introduction

Classical proteomics relies on the identification of proteins by mass spectrometry (MS). For simple mixtures containing modest numbers of proteins, the strategy of peptide mass fingerprinting (PMF) provides such identification. It consists of an enzymatic digestion of the proteins, followed by the determination of the masses of the resulting peptides. For any protein of known sequence, testing whether the expected set of peptide masses (*mass signature*) is present in the experimentally determined spectrum can be used to indicate the presence or absence of the protein. On complex mixtures, peptide masses tend to become ambiguous, and the PMF approach fails. This problem is typically overcome by fractionating the digested proteins by High Performance/Pressure Liquid Chromatography (HPLC), followed by mass spectrometry analysis. The mass analysis of each HPLC fraction is followed by the fragmentation of peptides isolated by MS and the analysis of the obtained fragments. As fragmentation patterns are predictable from protein sequence, the identification can then be achieved by matching experimentally obtained fragmentation patterns to predicted ones. The drawback of this approach is that only a limited number of peptides (typically, thousands) can be fragmented in a single experiment. Here, we demonstrate that proteins in complex mixtures can be uniquely identified using predicted retention time and without the need of any fragmentation experiments.

2 Methods

Our novel generalization of the PMF approach to HPLC data builds on recently developed accurate predictors of peptide *retention times* (*i. e.*, the time it takes the peptide to elute from the HPLC column)

[3]. We use the retention time to define an *ordered mass signature (OMS)*, in which the peptide masses expected for a single protein P are ordered according to their predicted retention time. We identify the best (partial) series of ions matching the OMS in an HPLC experiment by Dynamic Programming. This approach leads to an *OMS score* for each protein in an underlying database, with high scores corresponding to many matched peaks.

To evaluate the significance of a given OMS score, a decoy database has been generated by randomisation of the protein sequences [1]. The distribution of the scores has been analysed empirically using quantile regression [2]. We used data obtained by two of us (T.G. and D.G.) from the pathogenic bacteria *Francisella tularensis* substrain *novicida* and the corresponding protein sequences, available from NCBI (Accession Number NC_008601.1). Before scoring, each of the MS scans has been independently reduced to a list of monoisotopic, monocharged peaks following a procedure derived from [4].

3 Results

The empirical OMS score distribution on randomised proteins was obtained for limits corresponding to the increasing quantiles for the decoy proteins (see Table). The proteins with scores above these limits were considered present in the experiment with a number of false positive corresponding to the quantiles (20, 10, 5, and 1%).

To provide a coarse estimate of the Spurious Detection Rate, $\mathrm{SDR}=\frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}}$, we used false positive rates as given by the quantile bounds. Since the number of proteins in the sample (positive) was unknown, we conservatively chose the solution with the largest number of negatives (cf table 1).

Finally we used additional fragmentation data available for the dataset to obtain protein identifications for comparison. A Mascot search of these data with mass tolerance of 1.2 Da (peptides) and 0.6 Da (fragments) detected 413 proteins with the default settings (5%). The number of Mascot-detected proteins above each of the considered quantile limits is indicated on the last line of table 1.

quantile	0.80	0.90	0.95	0.99
real data	624 (36.3%)	458 (26.6%)	312 (18.2%)	154 (9.0%)
decoy data	334 (19.5%)	174 (10.2%)	87 (5.1%)	15 (0.9%)
SDR estimate	44%	31%	24%	10%
Mascot overlap	315	266	206	121

Table 1. Number and proportion of proteins above each quantile limit: Proteins (real data), Randomised proteins (decoy data), Estimated proportion of proteins above quantile limit explained by the negative model (SDR estimate), Intersection with proteins detected from fragmentation data (Mascot overlap).

References

[1] J.E. Elias and S.P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4:207-214, 2007.
[2] R. Koenker, quantreg: Quantile Regression, R package version 4.24, 2008.
[3] O.V. Krokhin, Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. *Anal Chem*, 78:7785-7795, 2006.
[4] R. Matthiesen, Extracting monoisotopic single-charge peaks from liquid chromatography-electrospray ionization-mass spectrometry. *Methods Mol Biol*, 367:37-48, 2007.

Multiple perturbation mapping of biological systems

Magali Michaut[1], Gary Bader[1]

[1] Terrence Donnelly CCBR, University of Toronto,
160, College Street, Toronto, Ontario, M5S 3E1, Canada
magali.michaut@utoronto.ca
gary.bader@utoronto.ca

Abstract: *The global aim is to develop a computational cell map that organizes all biological processes in yeast and their component interactions and molecules, and to use it to gain new biological insight. Specifically, we aim to use a multiple perturbation approach to map the yeast cell.*

Keywords: Genetic interactions, networks, cell map, multiple perturbation.

1 Aims

The global aim is to develop a computational cell map that organizes all biological processes in yeast and their component interactions and molecules, and to use it to gain new biological insights. Perturbations have been employed to assess the role of each gene in a specific cellular function in yeast, such as growth in normal conditions. Single perturbation often leads to little phenotypic effects because of compensation mechanisms. Thus multiple concomitant perturbations are needed to identify the contributions of most genes to specific functions. Genes that when perturbed together have an unexpectedly strong or weak effect on the cell are said to genetically interact. These interactions can be aggravating, alleviating or neutral. Aggravating interactions can reveal compensatory or parallel function, and alleviating interactions reveal order of action in the underlying biochemical network. This relationship between genetic and physical interaction networks allows us to organize genes into biological systems. By gathering large amounts of genetic interactions, we will determine a system view of the yeast cell. This map will then be read to understand how biological processes work, what is the function of a gene, and will provide insight into the cellular effects of disease-associated or engineered perturbations.

2 Methodology

2.1 Develop a hierarchical system map of the yeast cell

Genetic interactions are detected between two genes when the phenotypic consequence of perturbing both genes is different than expected given the phenotypes of each single gene mutant [1]. Synthetic lethality is a form of this relationship, occurring when the double mutant is inviable while each single mutant survives. Synthetic lethal interactions typically occur between genes in separate, but parallel pathways. Identifying significant numbers of genetic interactions between

gene sets reveals each set as a pathway or complex. We will work to show that synthetic genetic interactions map buffering relationships among biological pathways and summarize these relationships between known gene sets to develop and visualize a system level cell map. This will help delineate the boundaries of complexes and pathways involved in specific cellular phenotypes.

The first synthetic genetic array (SGA) data [2] enabled collection of qualitative genetic interactions based on the growth of the cell (*e.g.* growing, sick, lethal). Segre *et al.* computed synthetic genetic interactions using a model of yeast metabolism involving 890 genes [3], though this does not consider more than 5,000 other genes in yeast. The SGA technology has recently been extended to measure quantitative interactions based on level of cell growth measures by colony size. We will use novel quantitative SGA data and novel algorithms to define buffered protein complexes and pathways. The use of additional genetic interaction types that can be defined using these quantitative phenotypes will enable more specific predictions about the underlying physical interaction network. For instance, epistasis interactions link upstream to downstream genes. The novelty of this aim derives from the use of higher resolution synthetic genetic interactions.

2.2 Test the correctness of the cell map using orthogonal data

First, we will benefit from new genetic interaction data from high-resolution phenotypes derived from cell images. This technology uses a high-throughput microscope to analyze over 30,000 cell images per day. Analyzing these images generates over 300 phenotypic parameters for each cell imaged (e.g. spindle length, cell size). We will use these more specific phenotypes to gain new biological insight about the phenotypes measured.

Second, we will use chemogenomic assay information. The genome-wide collection of yeast gene deletion strains [1] has been used to generate genetic profiles of drug sensitivity and resistance. A proof-of-principle study has shown that the integration of chemical-genetic and genetic interactions data could give new insights into target pathways and proteins [4]. Chemical genomic assays have recently been performed on the yeast whole-genome heterozygous and homozygous deletion collections [5]. We will use these chemical-genetic profiles to define systems in the cell and test the correctness of the systems map.

Third, we will combine genetic interactions and protein-protein interactions which provide complementary information and enable more accurate system definition. Ulitsky *et al.* have combined these interactions to identify functional modules [6] but they have only considered physical interactions inside complexes. We aim to combine new quantitative genetic interaction data and different types of physical interactions including binary interactions.

References

[1] A. Tong *et al.*, Global mapping of the yeast genetic interaction network. Science, 303(5659):808-813, 2004.
[2] A. Tong *et al.*, Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science, 294(5550):2364-2368, 2001.
[3] D. Segré *et al.*, Modular epistasis in yeast metabolism. Nat. Genet., 37(1):77-83, 2005.
[4] A. Parsons *et al.*, Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. Nat. Biotech., 22(1):62-69, 2004.
[5] M. Hillenmeyer *et al.*, The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science, 320(5874):362-365, 2008.
[6] I. Ulitsky *et al.*, From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions. Mol. Sys. Biol., 4:209, 2008

Dynamic modelisation of transcriptional regulatory networks involved in yeast antifongal resistance

Jennifer BECQ[1], Sophie LEBRE[2], Frédéric DEVAUX[3] and Gaëlle LELANDAIS[1]

[1] Equipe DSIMB, INSERM UMR S665, Université Paris 7,
INTS, 6 rue Alexandre Cabanel, 75015 Paris, France
jennifer.becq@univ-paris-diderot.fr ;
gaelle.lelandais@univ-paris-diderot.fr
[2] Université de Strasbourg , LSIIT, CNRS UMR 7005
Pôle API, Bd Sébastien Brant - BP 10413 67412 Illkirch cedex, France
lebre@lsiit.U-Strasbg.Fr
[3] Laboratoire de Génétique Moléculaire, CNRS UMR 8541
Ecole Normale Supérieure, 46 rue d'Ulm, 75230 Paris cedex 05, France
devaux@biologie.ens.fr

Abstract: *For a cell to operate accurately, the expression of its genes must be precisely coordinated through regulatory systems. Among the numerous existing regulatory interactions, some rely on specific transcriptional modules driven by transcription factors. The yeast "Pleiotropic Drug Resistancce" (PDR) network is one such regulatory system, which we propose to caracterise over time. This network is involved in yeast resistance to the presence of a toxic component in the environment. The network described up to now is quite complex (figure 1); a unique transcription factor can regulate the expression of numerous target genes, and a single target gene can be regulated by different transcription factors. This description of the PDR network is a static picture of the relations between genes: it shows all the possible relations, and when the cell responds to physiological conditions it will use only a subset of these relations. In order to assess the dynamic of the PDR network, very resolutive analyses of the transcriptome by DNA microarrays were carried out at the "Laboratoire de Génétique Moléculaire" (CNRS UMR 8541, ENS Paris). The responses of three species of yeast (*Saccharomyces cerevisiae*, *Candida glabrata *and* Candida albicans*) to different toxic agents were studied: benomyl [1,2], fluphenazine [3], progesterone [4] and selenium [4]. Also, we recently developped a method to infer regulatory networks based on cinetic gene expression data [6]. The asset of this method is that it is be able to assign over time different structures of a network (*i.e. *the interactions between transcription factors and target genes) as well as the time intervals for which each network structure is effective. We are currently using this method to propose time-varying dynamical models of the transcriptional regulatory networks involved in the response to each of the chemical components cited earlier. We shall also add a evolutionary perspective to these regulatory networks caracterisations by carrying out comparative analyses between the different yeast species. Any finding of conserved or species-specific interactions should enlighten the mecanisms of antifungal resistance of the pathogens* C. albicans *and* C. glabrata.

Keywords: Transcriptome, réseaux de régulations, stress chimique, levures.

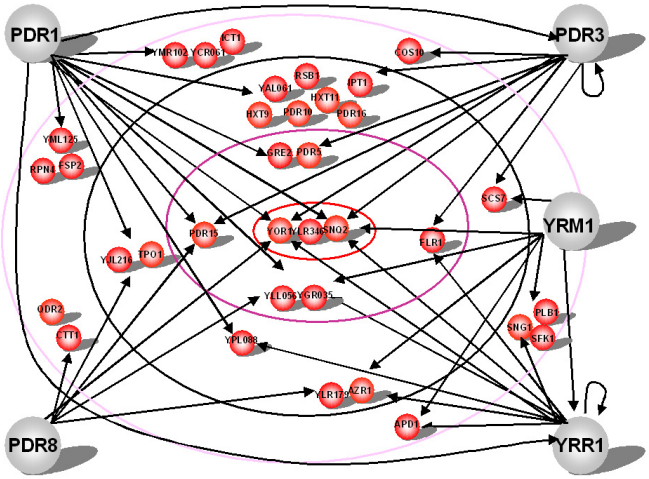


Figure 1. *S. cerevisiae* PDR network. Transcription factors and target genes are represented in grey and red respectively. These results were obtained by combining studies from the LGM (CNRS UMR 8541, ENS Paris).

References

[1] A. Lucau-Danila*, G. Lelandais*, Z. Kozovska, V. Tanty, T. Delaveau, F. Devaux and C. Jacq, The Early Expression of Yeast Genes Affected by Chemical Stress. Mol Cell Biol. 2005 Mar;25(5):1860-8.

[2] G. Lelandais, V. Tanty, C. Geneix, C. Etchebest, C. Jacq and Devaux F, Genome adaptation to chemical stress: clues from comparative transcriptomics in *Saccharomyces cerevisiae* and *Candida glabrata*. Genome Biol. 2008;9(11):R164. Epub 2008 Nov 24.

[3] V. Fardeau, G. Lelandais, A. Oldfield, H. Salin, S. Lemoine, M. Garcia, V. Tanty, S. Le Crom, C. Jacq, and F. Devaux, The Central Role of PDR1 in the Foundation of Yeast Drug Resistance. J Biol Chem. 2007 Feb 16;282(7):5063-74. Epub 2006 Dec 11.

[4] D. Banerjee, G. Lelandais, S. Shukla, G. Mukhopadhyay, C. Jacq, F. Devaux and R. Prasad, The Steroid Responses of Pathogenic and Non-Pathogenic Yeast Species Enlighten the Functioning and Evolution of Multidrug Resistance Transcriptional Networks. Eukaryot Cell. 2008 Jan;7(1):68-77. Epub 2007 Nov 9.

[5] H. Salin, V. Fardeau, E. Piccini, G. Lelandais, V. Tanty, S. Lemoine, C. Jacq and F. Devaux, The Steroid Structure and properties of transcriptional networks driving selenite stress response in yeasts. BMC Genomics. 2008 Jul 15;9:333.

[6] S. Lèbre, G. Lelandais, F. Devaux, Inferring changes in regulatory network structure from gene expression data. Submitted.

Posters

Factor VIII/von Willebrand Factor complex inhibits RANKL-induced osteoclastogenesis and controls cell survival

Marc BAUD'HUIN[1,2], Laurence DUPLOMB[1,2,3], Stéphane TELETCHEA[1,2,3], Céline CHARRIER[1,2], Mike MAILLASSON[4], Marc FOUASSIER[5], and Dominique HEYMANN[1,2,3]

[1] INSERM U957, ERI 7, Nantes, F-44035 France

[2] Université de Nantes, Nantes atlantique universités, Laboratoire de Physiopathologie de la Résorption Osseuse et Thérapie des Tumeurs Osseuses Primitives, EA3822, Nantes, F-44035, France

[3] CHU, Hōtel Dieu, Nantes, France

[4] INSERM U892 and IFR26- Ouest genopole, F-44035 France

[5] Centre Régional de Traitement de l'Hémophilie, Laboratoire d'Hématologie, CHU, Hōtel-Dieu, Nantes, France
stephane.teletchea@univ-nantes.fr

The molecular triad OPG/RANK/RANKL is the key regulator of bone biology and any change in the OPG/RANKL equilibrium leads to pathological conditions (1). The three proteins belong to the Tumor Necrosis Factor (TNF) family. Receptor Activator of Nuclear factor κB Ligand (RANKL) is mainly expressed by osteoblasts in bone micro environment and acts as a pro-resorption factor (2) . RANKL binds to its receptor RANK, expressed at the membrane of osteoclast precursors, to promote osteoclastic differentiation and maturation, this leads to bone resorption(3). Osteoprotegerin (OPG), also mainly produced by osteoblasts, is a soluble decoy receptor for RANKL which prevents the binding of RANKL to RANK, thus inhibiting osteoclastogenesis (4). Factor VIII (FVIII)/von Willebrand Factor (vWF) complex is involved in the coagulation cascade (5). Recently it has been described that vWF is physically associated with the anti-osteoclastic OPG, revealing its possible role in bone biology (6). The aim of this study was to determine the role of vWF on osteoclastogenesis and tumour cell survival.

To clarify the interactions between FVIII/vWF complex, RANKL and OPG, surface plasmon resonance and molecular modelling analysis were carried out. With the plasmon resonance experiments, we confirmed that OPG binds FVIII/vWF complex through the vWF. Interestingly, RANKL also binds to /vWF complex not to the recombinant FVIII or vWF alone, thus explaining the effect of only the FVIII/vWF complex on the RANKL signalling pathway.

To explore the putative binding mode between OPG and vWF, and since no experimental structure was available, we had to build the three-dimensional structural model of OPG using the crystallographic coordinates of the TRAIL/DR5 complex (7), and the model for human RANKL from the RANKL mouse structure (8). Since human and mouse RANKL proteins have a high sequence identity (90%), building and validating the model was straightforward. On the opposite, for the OPG model, its sequence identity with DR5 is only 27%. We used multiple alignments within the TNF superfamily for Mammalia to identify the conserved cystein-rich domains, characteristic of the extracellular regions in receptors. These domains were used to anchor the OPG sequence on the available DR5 structure (these domains are known to be involved in the OPG-RANKL binding site (9)). OPG and RANKL models were refined in Discovery Studio 2.1 (energy minimisations on loops and in regions in interaction in the complexes), each model quality was assessed using the Protein Health module to remove main chain and side chains disallowed regions. The OPG/vWF (A1 domain) complex was obtained by molecular overlay on the OPG/RANKL complex. Docking experiments (ZDOCK) were performed to improve the fit but did not allow to get a significantly better conformation than the result of the molecular overlay procedure.

From the OPG/vWF(A1) model, we observe that the OPG/A1 interaction surface is close to the OPG/RANKL one, but with a different binding mode (two major interaction sites for OPG/RANKL, one broader interaction site for OPG/vWF). The A1-recognition sites for other known agonists (GPIb, heparin) possess shared amino-acids with the predicted OPG binding site while antagonists are in contact with the opposite face of the A1 domain..

We also analysed the possible binding of FVIII/vWF complex to another member of the TNF superfamily: the pro-apoptotic cytokine TRAIL, known to bind OPG. The surface plasmon resonance analysis showed that FVIII/vWF complex or recombinant FVIII binds to TRAIL. TRAIL also binds to FVIII without affecting TRAIL/OPG interactions. This suggests a different binding mode for TRAIL between OPG and FVIII *in vitro*. We studied the *in vivo* effects of FVIII/vWF in a viability assay using human cancer cells sensitive to TRAIL. The FVIII/vWF complex abolished the OPG inhibitory activity on TRAIL-induced apoptosis. This apparent contradiction may come from an indirect shielding of the OPG/TRAIL binding site by the larger FVIII/vWF.

The present work is the first evidence of FVIII/vWF complex direct activity on osteoclasts differentiation and on induced cell apoptosis. We showed that FVIII/vWF complex inhibited RANKL-induced osteoclastogenesis in a dose-dependent manner. We were able to identify the putative binding mode from structural models derived from three-dimensional crystallographic structures, bioinformatics analysis and molecular modelling experiments.

These findings allows to embrace an enlarged knowledge on the proteins involved in physiological bone remodelling or in bone damages associated with severe haemophilia and cancer diseases.

Keywords: homology modelling, protein docking, osteoclastogenesis, TNF superfamily

Bibliography

1. Baud'huin M, Lamoureux F, Duplomb L, Rédini F, Heymann D. RANKL, RANK, osteoprotegerin: key partners of osteoimmunology and vascular diseases. Cellular and Molecular Life Sciences. 64(18):2334-2350, 2007.

2. Theoleyre S, Wittrant Y, Tat SK, Fortun Y, Redini F, Heymann D. The molecular triad OPG/RANK/RANKL: involvement in the orchestration of pathophysiological bone remodeling. Cytokine Growth Factor Rev. 15(6):457-475, 2004.

3. Burgess TL, Qian Y, Kaufman S, Ring BD, Van G, Capparelli C, Kelley M, Hsu H, Boyle WJ, Dunstan CR, Hu S, Lacey DL. The ligand for osteoprotegerin (OPGL) directly activates mature osteoclasts. J. Cell Biol. 145(3):527-538, 1999.

4. Simonet W, Lacey D, Dunstan C, Kelley M, Chang M, Lüthy R, Nguyen H, Wooden S, Bennett L, Boone T, Shimamoto G, DeRose M, Elliott R, Colombero A, Tan H, Trail G, Sullivan J, Davy E, Bucay N, Renshaw-Gegg L, Hughes T, Hill D, Pattison W, Campbell P, S S, er, Van G, Tarpley J, Derby P, Lee R, Boyle W. Osteoprotegerin: A Novel Secreted Protein Involved in the Regulation of Bone Density. Cell. 89(2):309-319, 1997.

5. Hollestelle MJ, Thinnes T, Crain K, Stiko A, Kruijt JK, van Berkel TJ, Loskutoff DJ, van Mourik JA. Tissue distribution of factor VIII gene expression in vivo--a closer look. Thromb. Haemost. 86(3):855-861, 2001.

6. Shabazi S, Lenting PJ, Fribourg C, Terraube V, Denis CV, Christophe OD. Characterization of the interaction between von Willebrand factor and osteoprotegerin. Journal of Thrombosis and Haemostasis. 5(9):1956-1962, 2007.

7. Mongkolsapaya J, Grimes JM, Chen N, Xu X, Stuart DI, Jones E, Screaton GR. Structure of the TRAIL-DR5 complex reveals mechanisms conferring specificity in apoptotic initiation. Nat Struct Mol Biol. 6(11):1048-1053, 1999.

8. Ito S, Wakabayashi K, Ubukata O, Hayashi S, Okada F, Hata T. Crystal Structure of the Extracellular Domain of Mouse RANK Ligand at 2.2-A Resolution. J. Biol. Chem. 277(8):6631-6636, 2002.

9. Cheng X, Kinosaki M, Takami M, Choi Y, Zhang H, Murali R. Disabling of Receptor Activator of Nuclear Factor-{kappa}B (RANK) Receptor Complex by Novel Osteoprotegerin-like Peptidomimetics Restores Bone Loss in Vivo. J. Biol. Chem. 279(9):8269-8277, 2004.

Bioinformatics contribution for the study of the regulatory network involved during cancer cell response to chemotherapy

Pierre-Yves Dupont[1], Dominique Loiseau[2], Daniel Morvan[3], Aicha Demidem[1], Georges Stepien[1]

[1] INRA U1019, Unit of Human Nutrition, St Genes Champanelle
pierre-yves.dupont@clermont.inra.fr
[2] Inserm U694, University of Angers, Angers
[3] University of Auvergne, Clermont-Ferrand

Abstract: *Most of cancer cells have a glycolytic activity. Our study links the constitutive glycolytic activity and gene regulation in transformed cells. One of these genes, ANT2, should allow cells to keep their mitochondrial integrity through maintenance of internal membrane potential gradient ($\Delta\Psi m$). In this study, we correlated biological results to a specific gene regulatory network adjusting cell requirements. We compared the response of control and cancer cells (from hepatocarcinoma and osteosarcoma) to an anticancer agent. Treatment effects were tested on global metabolite profiling by NMR spectroscopy. Over and underexpressed enzymes involved in disrupted metabolic pathways were deduced from NMR metabolite profiles. We developed an informatics pipeline to analyze the mechanisms of transcriptional regulation of genes encoding for selected enzymes. Our biological results showed an increased energy request to regenerate $\Delta\Psi m$ in cancer cells. This bioinformatics study allowed us to: 1 - construct specific sets of regulatory sequences (modules) in gene promoters; 2 - scan the whole human genome for these regulatory modules; 3 - identify human genes including such modules in their promoter sequence. The set of genes either selected from the NMR analysis or deduced from bioinformatics analysis was used to construct a regulatory network involved during cancer treatment.*

Keywords: Transcriptional regulation, Chemotherapy, Metabolic pathways, ANT2, Cancer.

1 Introduction

Cancer cells exhibit increased glycolysis and mainly depend on this metabolic pathway for the generation of ATP because mitochondrial ATP production is almost inactive. ANT2 imports ATP into mitochondria and thus should allow cells to maintain their mitochondrial integrity (by maintaining their $\Delta\Psi m$). The ability of tumor cells to satisfy ATP requirements may be a critical factor for their survival and we hypothesized that this ability should depend on their bioenergetic background. To investigate this hypothesis, we tested the capacity of tumor cells to shift from a glycolytic metabolism to a more oxidative one. We compared the response to an anticancer agent (CENU) of two transformed cell lines presenting different characteristics: a partially differentiated cell type (from hepatocarcinoma) and an undifferentiated one (from osteosarcoma). To this aim, we use a conventional anticancer agent provoking nuclear DNA damage and unknown mitochondrial effects. CENU interfere with cell

cycle and cell differentiation programs. Metabolomic analysis based on proton nuclear magnetic resonance spectroscopy of CENU treated cancer cells showed alteration of metabolic pathways, involving reprogramming of their glycolitic metabolism (TCA cycle).

2 Biological results

We examined mitochondrial functions after CENU treatment of the two cell lines. We mostly observed phospholipid derivative alterations. Mitochondrial integrity was evaluated by testing on mitochondrial respiration, $\Delta\Psi m$ maintenance and cell ATP production. We show that CENU has specific and early effects on mitochondria which could be compared with that of uncouplers, known to permeabilize the mitochondrial internal membrane and consequently to reduce the $\Delta\Psi m$, leading to an increased energy demand. The metabolic adaptation required to produce cellular energy was studied in two different transformed cell lines. We show that hepatocarcinoma cells, which have maintained differentiation properties, could switch to an oxidative metabolism with a mitochondrial ATP synthesis to respond to energy requirements. At the opposite, undifferentiated transformed cells (from osteosarcoma) with inefficient oxidative capacities could not maintain sufficient ATP production and were directed to cell death

3 Bioinformatics analysis and contribution

Within a web site, a bioinformatics pipeline was developed to identify human genes regulated by specific promoter sequence modules: combination of short regulatory sequences (called elements or matrices). Using Genomatix databases and tools, about 80 modules including 4 to 6 elements were constructed from the ANT2 gene promoter sequence and they were screened in the whole human genomic sequence. This analysis led us to select about ten specific regulatory modules with similar matrix combinations. A screening of the EnsEMBL gene database allowed us to identify about 20 genes or coding sequences downstream the selected modules. Interestingly, all characterized genes are involved in cancer cell metabolism. Finally, all the results where saved through an online database which allows to share data for all steps of this study. This software enables to retrieve much information on all selected genes from many public databases like Kegg, Unigene, Pubmed, Geo, etc. All this information is required to create models of metabolic pathways.

4 Conclusion

Taken together, we propose that mitochondrial oxidative background is an important clue of tumor cell fate in response to anticancer agents. It confers a survival advantage to more differentiated cells in response to chemotherapy. The set of genes either selected from the NMR analysis or deduced from bioinformatics analysis was used to propose a regulatory network involved during CENU treatment. One of the main features of this network is the coregulation of genes involved in cancer cell bioenergetics such as ANT2 and HIF1α. This bioinformatics study allowed us to identify similar regulatory sequences in their promoter.

PhEVER: Phylogeny and Evolution of Viruses and their Eukaryotic Relations

Leonor Palmeira[1,2,*], Simon Penel[1,*], Nicolas Girard[1], Vincent Lotteau[3], Christian Gautier[1,2] and Chantal Rabourdin-Combe[3]

[1] Université de Lyon, F-69000, Lyon
Université Lyon 1
CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive
F-69622, Villeurbanne, France.

[2] PRABI, Pôle Rhône-Alpes de Bioinformatique

[3] INSERM, U851, Lyon, F-69007, France
Universite de Lyon, Lyon, F-69003, France
Université Lyon 1, Lyon, F-69003, France
IFR128, Lyon, F-69007, France

{palmeira;penel;cgautier}@biomserv.univ-lyon1.fr
{vincent.lotteau;chantal.rabourdin}@inserm.fr

[*] These authors contributed equally to this work.

Abstract: *We present PhEVER, a database aiming at providing information for the analysis of co-evolution between viruses and their eukaryotic hosts. It was constructed from all viral genomes available and from the complete genomes of human and several insects vectors of pathogenic viruses. It groups sequences in clusters based on homology at both (1) the domain level and (2) the protein level and offers pre-computed alignments and phylogenies for each family, as well as an interface for domain visualisation and an integration of protein-protein interaction data. We focus on some interesting results obtained from the constructed clusters, and namely in the homologies between human and viral sequences.*

Keywords: Comparative genomics, phylogeny, co-evolution, lateral gene transfer, modules.

1 Introduction

Viruses are small particules of nucleic acids assembled in a proteic envelope which can develop a variety of types of interactions with the different organisms they co-exist with. How these complex interactions between several organisms are acquired and maintained throughout evolution is a question which remains open. Answering this question implies adressing questions on the exchange of genetic information between the interacting genomes such as: are some sequences more prone to horizontal transfer than others? and if so, are they related to specific metabolic pathways in the hosts? Determining if transfers are specific to families of viruses or if some global trends can be determined is also crucial for the understanding of the co-evolution between viruses and their host, in particular in the case of several intermediary hosts. We therefore namely focused in this study in detecting how genetic material is exchanged between the different interacting genomes and in determining how these exchanges affect both virulence and immunity mechanisms.

Viruses are usually mostly considered as pathogens responsible for a wide range of human pathologies, ranging from common cold to cancer, as it is now acknowledged that viruses are responsible for around 15 to 20% of human cancers. Hosts can be infected through air, direct contact, but also through vectors, namely insects. This is the case of highly virulent viruses such as the Flaviviruses responsible for dengue, Yellow fever or West Nile fever. The availability of a very large number of viruses, of an increasing number of pathogen vectors – *Aedes aegypti*, *Anopheles gambiae*, *Culex pipiens quinquefasciatus* – and of the human genome allows for a global approach on this matter.

2 State-of-the-art

In order to perform efficient comparative genomics analysis on the lasting interactions between viruses and their hosts, a first step is to be able to group related sequences together. This can, for instance, help detecting horizontal transfers of genetic material. However there is no available database integrating genomic data on both viruses and hosts. Most available viral databases are aimed at providing information on a specific viral family (HIV, influenza and Hepatitis C databases of the Los Alamos National Laboratory or VirusBanker for the Bunyaviridae). The only global database integrating families of homologous proteins between several viruses is the Viral Orthologous Cluster (VOCs) developped and maintained at the Viral Bioinformatics Resource Center. It however does not contain data on host sequences. We have therefore built a database focusing on homologies between viral and host sequences based both on the proteic level and on the domain level.

3 Results and Discussion

A complete set of non redundant viral sequences was obtained from RefSeq, and the complete genomes of human, *Aedes aegypti*, *Anopheles gambiae*, *Culex pipiens quinquefasciatus* were downlodaded from public databases. We have built a database of viral, human and insect vectors proteins using both (1) a similarity-based clustering algorithm which identifies homologies between complete proteic sequences and groups them into families and (2) a domain-based clustering algorithm which identifies homologies between segments of sequences and groups them into families. The similarity searches and clustering algorithms were adapted to fit the very divergent data. In order to integrate data for easy interpretation of evolutionary history, we pre-computed alignments of the sequences inside each family and inferred corresponding phylogenetic trees. An interface for domain visualisation as well as an integration of protein-protein interaction data offers an efficient tool for studying the impact of lateral gene transfer on the metabolism of host cells.

First of all, we obtain results regarding the homology between viral sequences and human sequences. We confirm the presence of retroviral oncogenes derived from cellular oncogenes in retrotranscribing RNA viruses – such as the Alpharetroviruses. We also find some typical retroviral oncogenes such as the serine/threonine-protein kinase family that present homologies with non-retrotranscribing viruses – such as the Phycodnaviridae – and which could be identified as putative oncogenes. We also present some results on the evolution by modularity notably on polyprotein-containing viruses such as the Flaviviridae. The strong homology of giant Mimivirus sequences, presenting a large number of similar domains with human sequences, is here reaffirmed as well as the uniqueness of Globuloviridae which seem to present no similarity with current available sequences.

Acknowledgements

This work is supported by the Région Rhône-Alpes.

ParameciumDB, a community model organism database built with the GMOD toolkit

Olivier Arnaiz, Jean Cohen, Linda Sperling

Centre de Génétique Moléculaire du CNRS
Avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex France
Olivier.Arnaiz@cgm.cnrs-gif.fr
Jean.Cohen@cgm.cnrs-gif.fr
Linda.Sperling@cgm.cnrs-gif.fr

Abstract: *Paramecium is a unicellular eukaryote that belongs to the ciliate phylum. The* Paramecium tetraurelia *genome is remarkably gene rich (~ 40 000 protein coding genes), the result of successive whole genome duplications (WGDs) in the Paramecium lineage. In order to build a model organism database capable of integrating the sequence and comparative genomic data from the genome project with genetic and other biological data from the community, we used the generic tools provided by the Generic Model Organism Database (GMOD) open source project: the Chado modular, ontology-based, database schema to store the data; the Turnkey web framework to render text; the Generic Genome Browser to render graphics; BioMart for the advanced query interface and Apollo for editing genome annotations.*

Keywords: Genomics, model organism, databases, whole genome duplication.

1 Introduction

The genome of *Paramecium tetraurelia,* recently sequenced at Genoscope [1], has been shaped by at least 3 whole genome duplications (WGDs), and a low rate of large-scale genome rearrangement facilitated reconstruction of the ancestral genome for each WGD and made it possible to establish synteny relationships between duplicated chromosomes. In order to build a *Paramecium* model organism database (http://paramecium.cgm.cnrs-gif.fr) [2] that integrates the data from the sequencing project with genetic and other biological data from the community, we used the generic tools provided by the Generic Model Organism Database (www.gmod.org) open source project.

2 Results

The data in ParameciumDB is stored in a relational database using the modular Chado schema developed by FlyBase [3], implemented under the open source RDBMS PostgreSQL. All of the data is typed with ontology terms, so that new classes of sequence-related data can be integrated

without any changes to the schema.

Type of data	Chado Module	Number of features
Sequence feature (all)	Sequence	1731294
gene	Sequence	40703
cDNA	Sequence	78110
comparative (all)	Sequence	739326
microarray probe signal	Mage	11658738
stock	Stock	1041
genotype + phenotype	Genetic + Phenotype	286 + 73
publication	Pub	2309

Table 1. Major types of data found in ParameciumDB (January 2009)

ParameciumDB text pages are rendered by the Turnkey/GMODWeb framework [4], which greatly accelerated design of the ParameciumDB web site. The Turnkey software first uses a code creation tool (Turnkey::Generate) to produce a model view controller (MVC)-based website given the database schema. A page-rendering module (Turnkey::Render) links the generated MVC code to an Apache mod_perl webserver. Finally, templates and cascading style sheets are used to customize the web pages.

ParameciumDB uses the Generic Genome Browser [5] for interactive genome browsing and for rendering graphic insets in Gene pages. BioMart [6] provides a data warehouse, an easily customized advanced query interface and web services. The Apollo Genome Editor [7], interfaced to the ParameciumDB Chado database using a JDBC direct read-write protocol, enables expert curation of the genome annotations by community members.

Acknowledgements

We acknowledge funding from the ACI IMPBio2004 #14 (L.S.) and the CNRS though the GDRE Paramecium Genomics (J.C).

References

[1] Aury, J.M. et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444: 171-8, 2006.

[2] Arnaiz, O., Cain, S., Cohen, J., and Sperling, L. ParameciumDB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data. Nucleic Acids Research 35: D439-44, 2007.

[3] Mungall, C.J., and Emmert, D.B. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics 23: i337-46, 2007.

[4] O'Connor, B.D., Day, A., Cain, S., Arnaiz, O., Sperling, L., and Stein, L.D. GMODWeb: a web framework for the Generic Model Organism Database. Genome Biology 9: R102, 2008.

[5] Stein, L.D. et al. The generic genome browser: a building block for a model organism system database. Genome Research 12: 1599-610, 2002.

[6] Kasprzyk, A. et al.EnsMart: a generic system for fast and flexible access to biological data. Genome research 14: 160-9, 2004.

[7] Lewis, S.E. et al. Apollo: a sequence annotation editor. Genome Biology 3: RESEARCH0082, 2002.

Whole genome evaluation of horizontal transfers for the pathogenic fungus *Aspergillus fumigatus*

Ludovic Mallet[1], Jennifer Becq[2] and Patrick Deschavanne[1]

[1] Molécules Thérapeutiques in silico (MTi) Inserm U973
Université Paris Diderot, Bâtiment Lamarck, $5^{ème}$ étage
5, rue Marie-Andrée Lagroua Weill-Halle 75205 Paris Cedex 13, France
ludovic.mallet@univ-paris-diderot.fr
patrick.deschavanne@univ-paris-diderot.fr
[2] Équipe DSIMB, Inserm UMR S665
Institut National de la Transfusion Sanguine
6, rue Alexandre Cabanel 75739 Paris Cedex 15, France
jennifer.becq@univ-paris-diderot.fr

Abstract: *Numerous cases of horizontal transfers (HT) have been described in eukaryote genomes, but no whole genome evaluation of HT has been carried out. The main reason being the lack of methods taking into account the intrinsic heterogeneity of these genomes. Here we propose to adapt a simple and tested method based on local variation of genomic signature to analyze the genome of the pathogenic fungus* Aspergillus fumigatus*. We detected about 1 Mb (3% of the genome) of DNA that exhibit atypical signatures. By analyzing the origin of these HTs by comparing their signatures to a home-made bank of species signatures, 3 major groups of donor species emerged: bacteria (40%), fungi (25%) and viruses (22%). It has to be noticed that though inter-domain exchanges are confirmed, we put in evidence only very few exchanges between eukaryotic kingdoms. In conclusion, we demonstrated that HT is quantitatively of a certain importance in eukaryote genomes though in a lower extent than in prokaryote genomes.*

Keywords: Horizontal transfers, eukaryote, whole genome, *Aspergillus fumigatus*.

1 Introduction

HT in eukaryotes are commonly detected by methods that spot incongruencies in phylogenetic trees. These methods work on a gene basis. However, the quantitative importance of HT in eukaryotes is poorly known though they were proposed to play a role as important as for prokaryotes, especially in fungi. But up to now, no whole genome evaluation of HT has been carried out for eukaryote genomes due to their complexity coming from non coding sequences, low complexity regions, isochores and fragmented genes. It was shown that variation of short oligonucleotide usage is moderate in some fungi genomes and that parametric methods based on this type of criterion could be applicable to them. Here we propose to adapt a simple and tested method based on local variation of genomic signature to analyze the genome of *Aspergillus fumigatus* which is an ubiquitous pathogenic fungus for a wide host range including Humans, ruminants and avians. This pathogen is implicated in different diseases from allergy to invasive aspergillosis eventually leading to the death of immunosuppressed patients.

2 Results

We used an HT detection method adapted from Dufraigne *et al* based on local variations of genomic signature of four-letters words. In a first step, all the centromeric and telomeric low complexity regions were removed from the genomic sequence. We then scanned the genome with sliding overlapping 5Kb-long windows, comparing the signature of each window to the signature of the whole genome in terms of Euclidian distance. We classified the windows in two classes on the basis of their genomic signature and their Euclidian distance to the whole genome. The class containing few windows showing high distances to the whole genome were excluded while the large class with homogenous signatures and small distance to the whole genome, called "host genome", was used to compute the HT detection treshold. One hundred and eighty nine distinct atypical regions were detected. They represent roughly 3% of the total genome (908 kb out of 29.4Mb). The average size of the atypical regions is 4.5 kb, ranging from 500 bp to 52.5 kb and they are widespread on all chromosomes. We searched the functions of the genes included in these atypical regions. One hundred and thirty four of these atypical regions contain 214 annotated genes, most of which exhibit homologous counterparts in Genbank with the exception of 21 ORFans. It can be noted that if 56% of them have a homolog only in fungi species, 19% exhibit homologs in other domains of life; for instance, 16 genes have homologs quite exclusively in bacteria. The functions of the transferred genes are mainly unknown. For the genes to which a putative function was inferred from annotation, half of them (23 out of 58) belong to central and intermediate metabolisms and none are involved in pathogenicity. Fifty-five atypical regions contain no annotated genes. BlastX and BlastN analyses detected gene relics in 24 (47%) of those regions. Besides some rRNA genes, detected by the method but supposedly not transferred, we found pseudo genes of nuclear or mitochondrial origin, transposons and plasmid parts. We compared the signatures of the 189 atypical regions to a hand-made bank of species signatures. For 117 regions, plausible donor species could be assigned. Three major groups of donors were identified: bacteria (40%), fungi (25%), and viruses (22%). Two groups are over-represented among the bacteria species: proteobacteria and actinobacteria. As a general trend, the origin of genes provided by the blastP analysis and the origin of the region proposed by the comparison of signatures agree. In conclusion, we demonstrated that HT is quantitatively of a certain importance in eukaryote genomes though in a lower extent than in prokaryote genomes. The global proportion of HT falls from the average 5.6% in tested prokaryotes to 3% in *A. fumigatus*. It remains to elucidate the biological mechanisms underlying those transfers and the importance of the biological role of the transferred genes.

Acknowledgements

JB and LM were supported by grants from the French Education and Research Ministry.

Lineage-specific pseudogenes identification through selective constraints analysis in the canine genome

Amaury Vaysse[1], Thomas Derrien[2], Catherine André[1], Francis Galibert[1], Christophe Hitte[1]

[1] Institut de Génétique et Développement de Rennes, UMR6061 CNRS, Université de Rennes1, 2 av. Pr. Léon Bernard, 35043 Rennes, France,
{amaury.vaysse,catherine.andre,francis.galibert,christophe.hitte}@univ-rennes1.fr

[2] Center for Genomic Regulation (CRG), Bioinformatics Program C/Dr aiguader, 88 08003 Barceona, Spain
thomas.derrien@crg.es

Abstract: *Lineage-specific gene-loss detection is an essential step to pinpoint the evolutionary forces that occur across species. Gene loss can derive from the process of pseudogenization that can occur rapidly under certain environmental circumstances or selection models. Pseudogenization may be detectable by the comparative analysis of nucleotide sequences between orthologous genes, through the ratio of replacement to silent nucleotide substitution rates (d_N/d_S). In this study, we used the d_N/d_S ratio to assess the level of selective constraints acting on 55 genes predicted as pseudogenes in the canine genome that are functional in primates and rodents. A strong relaxation of selective constraints was determined for pseudogenes with several in-frame mutations while no detectable relaxation of selective constraints was identified for pseudogenes with one in-frame mutation. This correlation may reflect different scenario of pseudogenization events in the dog genome. To perfom the analysis of d_N/d_S, we have developed OMEGA, a user-friendly web server that fully automates the analysis of selective constraints acting on genes across multiple lineages.*

Keywords: Selective constraints, pseudogenes, dog, evolution.

1 Introduction

Investigating the ratio of amino acid replacement (nonsynonymous, d_N) to silent (synonymous, d_S) substitution rate indicates selective constraints acting on a given protein-coding gene [1]. Therefore to assess the validity of pseudogene predictions at the nucleotide sequence level, the d_N/d_S ratio provides a proxy for the evolutionary constraints that occur on nucleotide substitution. The analysis of selective constraints using multiple species provides a mean to determine lineage-specific pseudogenes in a phylogenetic context. In such context, d_N/d_S ratio that show deviations from the expected rate of evolution in comparison to other species may reflect relaxation of constraints in a specific lineage. Here, we have investigated 55 canine pseudogenes through d_N/d_S analysis, for which we developed OMEGA, a web server that automates the analysis of selective constraints acting on genes across multiple lineages.

2 Results

2.1 Canine-specific pseudogenes characterization

Among genes that are functional in human, chimpanzee, rat and mouse species, we previously characterized 55 gene predictions containing ORF-disrupting mutations that suggested pseudogenization events in the dog genome [2]. Two subsets of pseudogenes were identified; a subset of pseudogenes (n=21) with accumulated mutations (mean=4.2) and a subset of pseudogenes with one mutation (n=34). To further characterize these subsets, we calculated the d_N/d_S ratio for each of the candidate pseudogene in comparison to their human functional orthologous gene from pairwise transcripts pair alignments. Assuming a constant mutation rate, the d_N/d_S ratio between dog pseudogenes and their human functional orthologs should theoretically relax towards 0.55 (average of 1.0 in the absence of selection and <0.1 for negative selection). For the pseudogenes with accumulated mutations, we calculated a median d_N/d_S of 0.50 indicating a considerable relaxation of selective constraints of the canine pseudogenes. For the pseudogenes with one mutation, a median d_N/d_S of 0.18 suggested no detectable differences in selective constraints between predicted canine pseudogenes and their human functional counterparts. These results further validated the 21 canine-specific pseudogene predictions with accumulated mutations. Interestingly, it raises the question whether pseudogene predictions with one mutation correspond to sequence artifacts or may have been inactivated much more recently in comparison to pseudogenes with several mutations.

2.2 A web server for automating the analysis of d_N/d_S

To fully automate the computation of d_N/d_S, we have developed OMEGA, a user-friendly web server. OMEGA can be used to analyze personal data for which it frees the user from the tedious task of computing codon-based alignement and d_N and d_S calculation. The server is also designed as a resource through which one can extract for a gene or groups of gene of interest pre-inserted d_N, d_S and values. To analyze personnal datasets, the server builds multiple sequence codon-based alignments, calculates the ratio of silent to replacement nucleotide substitution rates (d_N/d_S), provides features to calculate statistical analysis of the query set in comparison to benchmark set and returns a graphical display of the d_N/d_S distribution. In contrast to other tools [3], OMEGA is not limited to pairwise analyses; it is designed to analyze orthologous gene sets between several species ($n \geq 3$) enabling the assessment of selective constraint of each lineage in comparison to the rest of the phylogenetic tree. The server is designed to run in batch mode analysis with up to 100 sequences sets that can be submitted at the same time. OMEGA is available at: *http://dogs.genouest.org/OMEGA.html*

Acknowledgements

We thank the OUEST-genopole bioinformatics plate-form for hosting the web-server and for technical help.

References

[1] Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci, 13(5):555-556, 1997.
[2] T. Derrien, J. Thézé, A. Vaysse, C. André, E.A. Ostrander, F. Galibert and C. Hitte, Revisiting the missing protein-coding gene catalog of the domestic dog. BMC genomics, 10(1):62, 2009.
[3] M. Suyama, D. Torrents and P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res., 34:609-12, 2006.

CSPD: an *in silico* model for predicting carbonylated sites in proteins

Etienne Maisonneuve[1], Adrien Ducret[1], Pierre Khoueiry[1], Sabrina Lignon[2], Sonia Longhi[3], Emmanuel Talla[1], and Sam Dukan[1]

[1]Laboratoire de Chimie Bactérienne – Aix Marseille Université - UPR 9043-CNRS, 31 chemin Joseph Aiguier, 13402, Marseille Cedex 20, France
talla@ifr88.cnrs-mrs.fr

[2]Service de micro séquençage et de spectrométrie de masse – CNRS, 31 chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

[3]Architecture et Fonction des Macromolécules Biologiques, Aix Marseille Université - CNRS, UMR 6098, 163 Avenue de Luminy, Case 932, 13488 Marseille, Cedex 9

Keywords: metal-catalysed oxidation, carbonylated site and protein detection (CSPD), hot spot of carbonylation

The most important mechanism of oxidative damage in proteins is metal-catalysed oxidation (MCO). Carbonyl derivatives are formed by direct MCO attacks on the amino-acid side chains of proline (P), arginine (R), lysine (K) and threonine (T) residues. In recent years, several groups have been working on the identification of Carbonylated Sites (CS) within proteins from various organisms ([1], [2], [3], [4], [5]). These studies have led to two major observations: (*i*) sites are selectively carbonylated among all carbonylatable sites and (*ii*) CS are mainly located at the protein surface. Although these studies have contributed to shed light on the basis of protein carbonylation specificity, to date no rule have been found to predict sites more prone to carbonylation. In this study and with the use of mass spectrometry analysis, we identified carbonylated sites in oxidized BSA upon *in vitro* MCO, as well as in 23 proteins shown to be carbonylated from exponentially grown *Escherichia coli.* Moreover, we observed that the majority of carbonylated sites are located within an hot spot of carbonylation defined as RKPT-enriched regions within a particular environment. These observations led us to propose an *in silico* model that allows the efficient and accurate prediction of sites and proteins more prone to carbonylation in the *E. coli* proteome. Finally, our predictive model was extended to the detection of carbonylated proteins formed via direct MCO attacks in all organisms. Consequently, we are currently build a web tool called CSPD that will allow online predictions of carbonylated proteins (http://lcb.cnrs-mrs.fr/CSPD/). Our presentation will discussed (*i*) the motif rules governing carbonylation of proteins as well as the description of the CSPD model; (*ii*) analysis of the performance of the CSPD tool and (*iii*) comparative analysis of the predictive carbonylated proteins in prokaryotes and eukaryotes. This work comes from a narrow collaboration between biologists and bioinformatics teams and we hope that it will supply an new dimension to the identification of carbonylated proteins.

Acknowledgements

This work was supported by ACI Jeunes Chercheurs, ANR blanche ANR-05-BLAN-SPV005511.

References

[1] S. Lee, N. L. Young, P. A. Whetstone, S. M. Cheal, W. H. Benner, C. B. Lebrilla, and C. F. Meares, Method to site-specifically identify and quantitate carbonyl end products of protein oxidation using oxidation-dependent element coded affinity tags (O-ECAT) and nanoliquid chromatography Fourier transform mass spectrometry. J Proteome Res 5, 539-547, 2006.

[2] H. Mirzaei, H., and F. Regnier, Affinity chromatographic selection of carbonylated proteins followed by identification of oxidation sites using tandem mass spectrometry. Anal Chem 77, 2386-2392, 2005.

[3] H. Mirzaei, and F. Regnier, Creation of allotypic active sites during oxidative stress. J Proteome Res 5, 2159-2168, 2006.

[4] H. Mirzaei, and F. Regnier, Enrichment of carbonylated peptides using Girard P reagent and strong cation exchange chromatography. Anal Chem 78, 770-778, 2006.

[5] A. Temple, T. Y. Yen, and S. Gronert, Identification of specific protein carbonylation sites in model oxidations of human serum albumin. J Am Soc Mass Spectrom 17, 1172-1180, 2006.

Analyse comparée des contacts protéiques définis par distances ou par diagrammes de Voronoï

Jeremy Esque[1], Christophe Oguey[1], Alexandre G. de Brevern[2]

[1] CNRS UMR 8089, Laboratoire Physique Théorique et Modélisation (LPTM), Université Cergy Pontoise, Site de Saint-Martin, 2, avenue Adolphe Chauvin, Pontoise, 95302 Cergy-Pontoise cedex,France.
jesque@u-cergy.fr, oguey@u-cergy.fr
[2] INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Paris Diderot - Paris 7, INTS, 6, rue Alexandre Cabanel, 75739 Paris Cedex 15, France.
alexandre.debrevern@univ-paris-diderot.fr

Abstract: *Three-dimensional structures of proteins are the support of their biological functions. Their folds are stabilized by contacts between residues. Inner protein contacts are generally described through direct interactions between side-chain atoms, i.e. atomic proximity. Using Voronoï-Delaunay tessellation software, VLDP, we compare contact distributions as given by the classical distance method and the parameter-free diagrammatic approach.*

Keywords: Voronoï, Delaunay, protein contact, folding, inter-residue interaction.

1 Introduction

Les structures tridimensionnelles des protéines sont le support de leurs fonctions biologiques. Les protéines peuvent être décrites à l'aide des structures secondaires ou d'alphabets structuraux [1]. Les interactions entre résidus sont essentielles pour le repliement des protéines et la stabilisation de leur structure. Entre autres, ces interactions sont assurées par des liaisons covalentes, dépendantes directement de la séquence en acides aminés. Des liaisons non covalentes, plus faibles, sont également impliquées dans l'édifice. Par exemple, les liaisons hydrogènes jouent un rôle majeur dans le repliement et la stabilisation des structures secondaires, hélices α et feuillets β. A ce titre, les interactions entre résidus font l'objet de nombreuses recherches.
En général, les contacts entre résidus sont déterminés en appliquant un seuil fixe de distance, le plus souvent entre Cα. Lors d'une précédente étude, nous avons montré que, selon le type de distance utilisé, le nombre moyen de contacts était très différent. De plus, il n'impliquait souvent pas les mêmes paires de sites [2]. Les diagrammes de Delaunay ou de Voronoï sont une alternative robuste à cette définition de contact, libre de tout seuil défini *a priori* [3]. La construction de Delaunay/Voronoï a été utilisée pour étudier le *packing*, les structures secondaires, etc [4]. L'objectif de cette étude est de comparer les contacts définis par l'approche classique des seuils et ceux recensés entre les cellules de Voronoï.

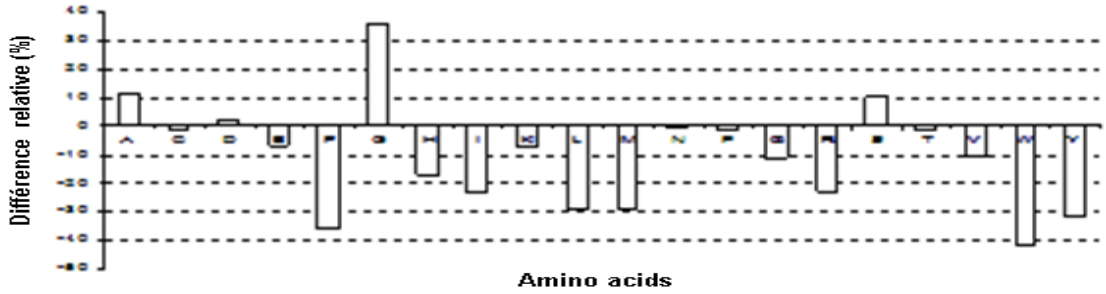


Figure 1. Différence (%) entre nombre moyen de contacts suivant les 2 méthodes

2 Définitions des contacts et tessellations

Classiquement, les contacts sont définis par un seuil de distance ; le plus souvent les distances sont mesurées entre Cα avec une valeur seuil de 8 Å [2]. En tessellation de Voronoï, les contacts sont les facettes séparant des cellules entourant des sites appartenant à des résidus différents. Les diagrammes de Voronoï-Delaunay ont été calculés avec le programme VLDP, développé au LPTM, à partir des coordonnées de tous les atomes ; la protéine est placée dans une boite d'eau équilibrée, une solution au problème de cellules du bord [5]. Ces deux points diffèrent d'analyses telles que [6]. L'algorithme VLDP procède par insertion séquentielle des sites, selon une méthode incrémentale à la fois robuste et efficace. VLDP contient plusieurs modules : calcul des aires et des volumes, composantes connexes, matrice de contacts, etc.

3 Principe de l'analyse et objectif de l'étude.

Nos statistiques portent sur un ensemble de structures protéiques, composé de 357 monomères, sélectionné sur les critères suivants: moins de 25% d'identité entre séquences; au moins 99% de résidus complets; compatibles avec les logiciels d'analyses dont nous disposons. Lors de l'analyse, les contacts à courtes distances dans la séquence (à moins de 4 résidus d'intervalle) sont éliminés, car ils ne sont pas informatifs. En utilisant les fréquences relatives de contacts [2], nous analysons les couples d'acides aminés sur- et sous-représentés, susceptibles de jouer un rôle privilégié au sein des structures protéiques. Selon nos premiers résultats, la méthode de seuil sous-estime fréquemment le nombre moyen de contacts (Figure 1). Des analyses en fonction de la taille, des structures secondaires des protéines ont également été réalisées.

Références

[1] B. Offmann, M. Tyagi and A.G. de Brevern, Local Protein Structures, Current Bioinf., 3:165-202, 2007.
[2] G. Faure, A. Bornot and A.G. de Brevern, Protein contacts, inter-residue interactions and side-chain modelling, Biochimie, 90:626-639, 2008.
[3] GF. Voronoi, Nouvelles applications des paramètres continus à la théorie de formes quadratiques, J angew Reine Math, 134 :198-287, 1908.
[4] A. Poupon, Voronoi and Voronoi-related tessellations in studies of protein structure and interaction, Cur Opinion in Struct Biol, 14:233-241, 2004.
[5] J.F. Sadoc, R. Jullien and N. Rivier, The Laguerre polyhedral decomposition: application to protein folds, Eur. Phys. J. B, 33:355-363, 2003.
[6] C.H. da Silveira, D.E. Pires DE, R.C. Minardi, C.J. Veloso, J.C.Lopes, et al, A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts, Proteins, 74(3):727-743, 2008.

Modeling and stability analysis of interconnected regulatory cycles

Mahsa Behzadi, Mireille Regnier, Laurent Schwartz, Jean-Marc Steyaert

Bioinformatics group, LIX, Ecole Polytechnique, Palaiseau, 91128, France
behzadim@lix.polytechnique.fr
mireille.regnier@inria.fr
laurent.schwartz@polytechnique.edu
steyaert@lix.polytechnique.fr

Abstract: *Our aim is to build a generic framework with which one could simulate the behavior of complex systems of interconnected regulatory cycles.For the simulation of a biological system we use the traditional reaction-rate approach by means of equations describing the system.Once constructed the model, we aim to study the various modes of the cell behaviour according to the concentrations of relevent enzymes in enzymatic reactions.Until now we have constructed a model for the central part of the system of the GlyceroPhosphoLipid metabolism in the human cell.We currently use this approach to study the stability analysis of a complex metabolic network containing several interconnected regulatory cycles.*

Keywords: ordinary differential equations, enzymatic reactions, stability analysis,cycle oscillations,equilibria.

1 Introduction

Biochemical reactions are continually taking place in all living organisms. The complexity of biochemical and biological processes is such that the development of computer models is often essential in trying to understand the phenomenon under consideration. Our aim is to build a generic framework with which one could simulate the behavior of complex systems of interconnected regulatory cycles.

2 Methods

For the simulation of a biological system we use the traditional reaction-rate approach by means of equations describing the system. In this approach, chemical reactions are modeled by ordinary differential equations (ODEs) representing the variations of the concentrations of the substances. In each of the differential equations we express the kinetics of one reactant as a sum of fractional terms for enzymatic reactions and non-fractional terms for simple reactions.

3 Analyses

Once constructed the model, we aim to study the various modes of the cell behaviour according to the concentrations of relevent enzymes in enzymatic reactions. Since stable and unstable equilibria

play different roles in the dynamics of a system, it is useful and important to be able to classify equilibrium points based on their stability, and this is what we are able to do by simulation and also by mathematical study. By stability analysis, first given an equilibrium we can determine if it is a stable point or not; furthurmore through a mathematical study based on differential equations we are able to find regions which correspond to stable steady-state behavior or cycle oscillations of the model by changing one or several parameters. This stability could be local or global. In some special cases we are able to prove mathematically that a stability is global.

4 Results and Models

As a first try we have constructed a model for the central part of the system of the GlyceroPhosphoLipid metabolism in the human cell. The Phospholipid metabolism has attracted the attention of many researches in cancer studies and they think the ability to follow phospholipid metabolism is of paramount importance in many circumstances in which cell survival and cell proliferation are of concern for example in neurological disorders and cancer . Thus there was an important need to develope a model for these biosynthesis, and that is the reason we tried to find a model for the important part of the system of GlyceroPhospholilid metabolism in the human cell. The model comprises enzymatic reactions of PhosphatidyleEthanolamine (PtdEth) and the PhosphatidylCholine (PtdCho)[**?**][**?**].Our analysis concerns twenty-four biochemical reactions. Given the values of metabolite concentrations (Ci) which were observed experimentally we have managed to find the appropriate parameter values (Pi) which allow us to completely describe the system with a set of ordinary differential equations (ODE). Our analysis of this model demonstrates that, with these parameter values, the system has a stable solution. Moreover, we investigatedthe possibility that a change in parameter values could give an unstable or oscillating solution. For that purpose we studied the system mathematically in a large range of values and we prove that the solution is always stable and without oscillations regardless the parameter values of the system.

We have also applied our method to the cell division cycle model; welknown interactions of proteins cdc2 and cyclin. A mathematical model was already constructed by Joun Tyson [**?**], who used numerical integration (carried out by using Gear's algorithm) for simulation and stability analysis of model. We studied this system of interactions and using our approach based on the analysis of the eigenvalues of the linearized system we confirmed the nature of the results for the same parameter values.

We currently use this approach to study the stability analysis of a complex metabolic network containing several interconnected regulatory cycles such as Glycolysis, Krebs cycle, Phospholipids pathway and Amino acids.

References

[1] Henry, S. A., and Patton-Vogt, J. L. (1998) *Prog. Nucleic Acids Res. Mol. Biol.* 61, 133-179

[2] R.Sundler,B.Akesson,*Biochem.J*.146(1975)309–315.

[3] J.Tyson,(1991) *Cell Biology*, Vol 88. pp. 7328–7332.

Co-evolution of blocks of residues and sectors in protein structures

L. Dib[1], A. Carbone[2]

Génomique Analytique, Laboratoire de Génomique des Microorganismes, FRE3214 CNRS-UPMC
15 Rue de l'Ecole de Médecine, 75005, Paris, France and Département d'Informatique, Université Pierre et Marie Curie-Paris6

[1] linda.dib@gmail.com
[2] Alessandra.Carbone@lip6.fr

Evolutionarily conserved networks of residues have been demonstrated to mediate allosteric communication in proteins involved in cellular signaling, the process by which signals originating at one site in a protein propagate reliably to affect distant functional sites. The general principles of protein structure that underlie this process remain unknown. In a seminal paper Ranganathan described a sequence-based statistical method for quantitatively mapping the global network of amino acid interactions in a protein [10,12]. The method reveals a surprisingly simple architecture for amino acid interactions in each protein family: a small subset of residues forms physically connected networks that link distant functional sites in the tertiary structure. The evolutionarily conserved sparse networks of amino acid interactions are proposed as representative structural motifs for allosteric communication in proteins. Investigating further Ranganathan approach, a new method, based on a fine combinatorial analysis of phylogenetic trees associated to a protein family has been developed to reconstruct networks of co-evolved residues from sequence analysis [3]. Various other approaches for identifying covariant amino acid pairs in protein sequences have been proposed [11,1,8,5]. Gloor [6] and Travers [7] methods have been designed to ensure that no detected signal is induced by phylogenetic side-effects due to an underrepresentation of sequences. Yeang [13] studied co-evolution in a domain and between domains.

We propose a method to detect co-evolved blocks of residues, numerically rank them depending on their level of co-evolution, and clusterize them to obtain networks of co-evolved blocks. In contrast to Ranganathan method, or to the various methods proposed in the literature, where coevolution of alignment columns is analyzed through a comparison of their residue distributions, our main focus is on groups of successive positions in the aligned homologous proteins, called *blocks*. In short, we compare the information content of pairs of blocks by looking at the way they evolve, possibly accepting exceptions. We define an appropriate score of co-evolution between pairs of blocks and develop a methodology to extract functional, structural and mechanical signals for protein families from co-evolved blocks. These signals will correspond to networks of co-evolved residues.

Given an alignment of n sequences, we consider groups of m consecutive positions in the alignment, where $m \geq 1$. For each group of consecutive positions and for each sequence in the alignment, there is a uniquely identified word that belongs to the group of positions and appears as a subword in the sequence. We look at the set of n words associated to a group of positions and study the combinatorial properties of the distribution of words in the group. Depending on these properties we shall say that a group is a block or not. Intuitively, we look at the space of all words of length m and check the variability of the words in a group. This is done by varying a parameter that accepts errors in words, and that allows words to be different but eventually partly conserved in several sequences. There are n different dimensions that are used to evaluate the space of words, and they correspond to the number of "errors" or "exceptions" that we want to accept. Dimension 0 is the most restrictive one and allows for no error, while dimension n allows for errors in all sequences. For each dimension, we evaluate pairs of groups of positions with a score of co-evolution and predict co-evolved blocks.

We report a detailed analysis realized on the MukB domain family. This family is highly conserved with an average identity of 79%. Although very conserved, our co-evolution method discovers, at dimension 0, only 3 blocks: the block from position 35 to 42, and residues at positions 144 and 162. The block from position 35 to 42 corresponds to the known motif $[AG](X)_4GK[ST]$, called the Walker-A motif, and reported in [9,2] for the putative G loop in MukB domain. The method proposed in [9] has been validated on this motif and a predicted 44 amino acids to be functionally important for the MukB domain family, where only 3 out of them are part of the G-walker motif. In contrast, our method selects 10 positions, 8 of which form the Walker-A motif. This finding confirms that by looking at co-evolved blocks provides fine predictions at a high level of specificity. Our co-evolution method discovers also two groups of highly co-evolved residues that surround this motif, both identified at dimension 1. They form two separated networks of blocks, made by residues that did not co-evolve with each other, and that in three dimensions, surround the Walker-A motif, like in a sandwich, with an upper layer and a lower layer. We call them sectors. This notion is novel and does not coincide with the one of a domain.
We consider a second example, the guanido phosphotransferase, C-terminal catalytic domain (PF00217 in PFAM:`1bg0`) analyzed in [4] where conserved positions that are involved in ligand binding sites are detected in addition to enzyme active sites. In contrast, we discover three independent networks at dimension 1, located just behind conserved hot spots. The networks are known to maintain the binding cavity of the enzyme active sites.
A large scale analysis of the performance of our method is underway.

Keywords: Coevolution, phylogeny, allosteric function.

References

[1] C. Ane, J. G. Burleigh, M. M. McMahon and M. J. Sanderson, Covarion structure in plastid genome evolution: a new statistical test. *Mol. Biol. Evol.*, 22:914-924, 2004.
[2] A. Armon,D. Graur, and N. Ben-Tal, Consurf: an algorithmic tool for the identification of functional regions by surface mapping of phylogenetic information.*J. Mol. Biol.*, 307, 447-463, 2001.
[3] J. Baussand, A. Carbone, A combinatorial approach to detect co-evolved amino-acid networks in protein families with variable divergence. *Submitted manuscript*, 2009.
[4] G. Cheng, B. Qian, R. Samudrala, D. Baker, Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Research*, 33:5861-5867, 2005.
[5] J. Dutheil, T. Pupko, A. Jean-Marie and N. Galtier, A model-based approach for detecting co-evolving positions in a molecule. *Mol. Biol. Evol.*, 22:1919-1928, 2005.
[6] SD Dunn, LM Wahl, GB Gloor, Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24:333-340, 2008.
[7] MA Fares, SAA Travers, A Novel Method for Detecting Intramolecular Coevolution: Adding a Further Dimension to Selective Constraints Analyses. *Genetics*, 173:9-23, 2006.
[8] N Galtier, Sampling properties of the bootstrap support inmolecular phylogeny: influence of nonindependence among sites. *Syst. Biol.* , 53:38-46, 2004.
[9] CA Innis, siteFiNDER—3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Research*, Vol. 35, 2007.
[10] SW Lockless, R Ranganathan, Evolutionary conserved pathways of energetic connectivity in protein families. *Science*, 286:295-299, 1999.
[11] DD Pollock, WR Taylor, Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng.*, 10:647-657, 1997.
[12] GM Suel, SW Lockless, MA Wall, R Ranganathan, Evolutionary conserved networks of residues mediate allosteric communication in proteins. *Nature Struct. Biol.*, 23:59-69, 2003.
[13] CH Yeang, D Haussler, Detecting coevolution in and among proteins domains. *PLoS Comp. Biol.*, 3:2122-2134, 2007.

Chromosome organization in *Buchnera*: a dynamic active structure involved in gene expression regulation

Lilia Brinza[1], Federica Calevro[1,3], José Viñuelas[4], Christian Gautier[2,3], Hubert Charles[1,3]

[1] UMR203 Biologie Fonctionnelle Insectes et Interactions (BF2I), IFR41, INRA, INSA-Lyon, Université de Lyon, F-69621 Villeurbanne, FRANCE
[2] Université de Lyon; Université Lyon 1; CNRS; UMR 5558; Laboratoire de Biométrie - Biologie Evolutive, Bâtiment Gregor Mendel, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne, France
[3] BAMBOO, INRIA Rhône-Alpes, France
[4] Université de Lyon, Université de Lyon 1, CNRS, UMR5534, Centre de Génétique Moléculaire et Cellulaire, F-69622 Villeurbanne, France

Abstract: *Genomic studies on bacteria have shown the existence of chromosomal organization. Moreover, transcriptomic analyses have demonstrated that, in free-living bacteria, gene transcription levels and chromosomal organization are mutually influenced. We analysed chromosomal organization structures likely to modulate gene expression in the highly reduced genome of Buchnera aphidicola, the primary endosymbiont of the aphids.*

Keywords: *Buchnera aphidicola*, transcriptional regulation, DNA topology, chromosome organization, intracellular symbiosis.

Most bacterial chromosomes consist of a single closed-circular DNA molecule folded into a compact and dynamic structure called the nucleoid. Variations of chromosome 3D-structure act as a global regulatory factor of gene expression. More particularly, the modulation of genome architecture is admitted to belong to the class of mechanisms allowing genome-wide transcriptional profile variation in response to environmental changes [1].

We searched for evidences of spatial organization of the chromosome in an extremely intriguing bacterial model: *Buchnera aphidicola*. Associated with most agricultural pest aphids and being partly responsible for their harmfulness, *Buchnera* are one of the most studied intracellular symbiotic bacteria of insects. Their genomes present all the characteristics of intracellular bacteria: (1) small size of 400 - 600 kb depending on aphid species, (2) highly biased base composition towards A and T and (3) high evolutionary rate due to the isolation of *Buchnera* populations within the host cells combined with the drastic bottlenecks that occur in the population dynamics of the bacteria during their transmission to the aphid progeny.

A crucial stage for the symbiosis comprehension passes through the understanding of symbiont gene expression regulation, and yet little is known about the transcriptional regulation capabilities of the bacteria. Given the “poor” catalogue of transcriptional factors it was suggested that the bacteria are no longer able to regulate their gene expression. Also, *Buchnera* conserved target genes for regulatory proteins absent in their genome. Nevertheless, recent works using a dedicated microarray showed that *Buchnera* respond specifically to some

nutritional stresses imposed to their host [2].

The aim of this work was to study potential structural units of the *Buchnera* chromosome and the impact they could have on the gene expression regulation. For this purpose, we analysed the potential structural domains of the chromosome at different scales and the main contributor proteins for the organization and maintenance of these domains. Our study brings evidences for (1) the existence of structural chromosomal units at several scales, (2) a functionally complete set of proteins essential for nucleoid organization and (3) the tight interdependence between these proteins, the chromosome architecture and the gene expression profile.

Basic genomic structural elements in bacteria participating to the chromosome organization are transcription units (operons). A transcription unit contains one or several adjacent genes transcribed as a single mRNA. Thus, genes belonging to the same transcriptional unit display strong correlated transcription levels. A first annotation of *Buchnera* transcription units was available in BioCyc (http://biocyc.org/). We found that some of these transcriptional units were not consistent with the analysis of the gene expression profile. Thus, we decided to re-annotate the transcription units of *Buchnera* taking into account gene expression levels, gene order conservation, sequence features of *Buchnera,* like Rho-independent terminators inferred withTransTermHP [3] and specific intergenic distances. We tested this new annotation with microarray gene expression data.

A higher level of structural units in bacterial genomes is represented by the organization in topological domains (~10kbp in *E. coli*). These structures are mainly organized and maintained by Nucleoid Associated Proteins (NAPs). Analysis of the NAP set of *Buchnera* pointed out that, despite genome reduction, the bacteria retains the most important members of the group (IHF, H-NS, Fis, DnaA and HU). Our bioinformatics analysis of these proteins confirms a strong conservation of structural domains and 3D structure. Moreover, key amino acids (their mutation compromises NAPs function in *E. coli*) are also well conserved. As NAPs must frequently bind DNA (every 10 kbp) to form dynamic inter-domains barriers, they rather recognize specific topologic structures (i.e., bend DNA) than specific motive binding sites.

Combination of topological domains and DNA fold are at the origin of a third class of larger structural units in bacterial chromosomes. Previous results of our team pointed out a periodic transcriptional pattern that supports the existence of these kinds of structures in *Buchnera* [4]. We completed this work by using a more realistic distance on the chromosome (*i.e.*, physical distance (bp), instead of the "gene number" distance).

Our work brings several evidences that *Buchnera* chromosome is a functional structure probably playing an active role in gene expression regulation. This kind of regulation was often neglected in free-living bacteria but might be central in shrunken genomes of endosymbionts. A short-term perspective of our work will be to inactivate *in vivo* specific NAP proteins of *Buchnera* and analyse the induced modifications within the gene expression profile of the symbiotic bacteria.

References

[1] Crozat, E., et al., Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. Genetics, 169(2): 523-532, 2005.

[2] Reymond, N., et al., Different levels of transcriptional regulation due to trophic constraints in the reduced genome of *Buchnera aphidicola APS*. Appl. Environ. Microbiol., 72(12): 7760-7766, 2006.

[3] Kingsford, C., K. Ayanbule, and S. Salzberg, Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biology, 8, 2007.

[4] Vinuelas, J., et al., Conservation of the links between gene transcription and chromosomal organization in the highly reduced genome of *Buchnera aphidicola*. BMC Genomics, 2007. 8(1): 143.

A Study of Genomic Rearrangements in Maize Mitochondrial Genomes

Aude Darracq[1,2], Jean-Stéphane Varré[2], Pascal Touzet[1]

[1] GEPV, UMR CNRS 8016, USTL
aude.darracq@ed.univ-lille1.fr, pascal.touzet@univ-lille1.fr
[2] LIFL, UMR CNRS 8022, USTL, INRIA Lille - Nord Europe
jean-stephane.varre@lifl.fr

Abstract: *Even though plant and animal mitochondrial (mt) genomes share a common ancestor, the plant mitochondrial genome exhibits peculiar features when compared with its animal counterpart: a large variation in size (200-800 kb in plants, 15-16 kb in animals) and in gene order due to intra-genomic recombination. This variation observable at the intra- and inter-species levels is caused by active recombination sequences, most likely acquired by the ancestral plant lineage. We compared eight maize mitogenomes and proposed a rearrangement phylogeny. Despite the fact that those genomes are highly rearranged and thought difficult to analyze, our observation led us to propose a phylogenetic tree combining inversions and tandem duplication events.*

Keywords: rearrangements, tandem duplication, mitochondrial genome, maize

1 Materials and methods

Dataset. We used eight *Zea mays* mitochondrial genomes whose sequences are available from GenBank [1,2]. Five of them are from *Zea mays* ssp. *mays*: two fertile cytotypes *NA* and *NB* , and three cytoplasmic male sterile (CMS): *CMS-C* ,*CMS-S* and *CMS-T*. The three others are *Zea mays* ssp. *parviglumis*, *Zea luxurians* and *Zea perennis*. The two last ones serve as outgroups. The length varies between 535,825bp and 739,719bp. The difference in length is mostly due to large duplicated parts found in 3 genomes (*NA*, *CMS-C* and *Zea mays* ssp. *parviglumis*).

Methods. We extracted annotated protein coding genes, tRNA, rRNA and ORFs from the genomes, called *genes*. We then compared genes in order to determine othologous and paralogous relationships using the reciprocical best hit principle. We numbered each of the 87 genes and obtained a sequence of numbers for each genome, with some duplicated numbers. Notice that duplicated genes are not always the same ones in each genome, and a duplicated gene is not necessarily duplicated in all genomes. Rearrangement phylogenetic analyses were performed using GRIMM [3] with Neighbor-Joining or MGR [6].

2 Results and discussion

Difficulty to take duplicates into account. The common way to deal with duplicates is i) to follow the exemplar model where exactly one copy of each gene is conserved or ii) to follow the maximal matching model where a maximum number of copies are conserved. In the latter case, one has to

determine orthologous relationships but as the duplicated genes are not the same in each genome, such a method is not suitable. With our data, the number of trees obtained with the exemplar model is more than 16 millions which makes this method unusable.

Model of tandem duplication with partial losses. When looking the gene sequences, we can see that duplicated genes are often grouped together in tandem (see for example *Zea mays* ssp. *parviglumis*, Figure 1 left, last row). Nevertheless, the set of genes is not exactly the same between the duplicated parts. This suggests that the set of genes involved into the duplication may have evolved by deletion of some of them after the duplication event. Such a *tandem duplication with partial losses* (TDPL) event explain the majority of the duplicated parts of the eight genomes. A similar event has been highlighted in animal mitogenomes [5]. Therefore we assume a common way of evolution between plant and animal mitogenomes due to their supposed common origin. Other duplicated genes are involved into a different context and thus we can easily distinguish orthologous from paralogous.

Our method. We used the following scheme: i) identify TDPLs and collapse them in order to keep one copy of each gene, ii) distinguish between paralogous and orthologous for remainded duplicated genes, iii) apply usual rearrangement algorithms (no duplicates remains), iv) expand the previously collapsed TDPL. Figure 1 gives an example explaining the evolution of the maize mitogenomes.

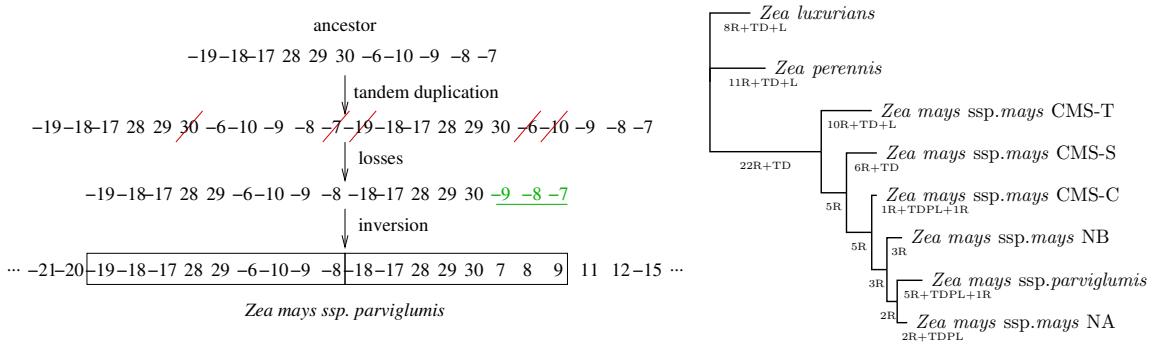


Figure 1. A scenario involving a TDPL for *Zea mays parviglumis* and a rearrangement phylogeny (inversions (R), duplications (TD), losses (L) and TDPLs).

3 Acknowledgments

Grants from the PPF Bioinformatique of the University of Lille 1 and from ANR (Jeunes Chercheurs).

References

[1] J. O. Allen *et al.* Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. *Genetics*, 177:1173–1192. 2007.
[2] J. O. Allen *et al.* The complete mitochondrial genomes of five close relatives of maize. Unpublished.
[3] G. Bourque and P. A. Pevzner. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Research*, 12(1):26–36, 2002.
[4] A. C. Darling, B. Mau, F. R. Blatter and N. T Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7):1394–1403. 2004.
[5] D. San Mauro, D. Gower, R. Zardoya, M. Wilkinson. A Hotspot of Gene Order Rearrangement by Tandem Duplication and Random Loss in the Vertebrate Mitochondrial Genome *Molecular Biology and Evolution*, 23(1):227–234, 2006
[6] G. Tesler. GRIMM: genome rearrangements web server. *Bioinformatics*, 18(3):492–493. 2002.

Finding miRNAs homologs in genomes with no learning

Anthony Mathelier[1,2], Alessandra Carbone[1,3]

[1] Analytical Genomics, FRE3214 CNRS
91 boulevard de l'Hôpital 75013 Paris
[2] anthony.mathelier@gmail.com
[3] alessandra.carbone@lip6.fr

Abstract: *MicroRNAs (miRNAs) are a class of endogenes of* $18-25$ *nucleotides (nt) in length which are derived from a precursor (pre-miRNA) with a characteristic hairpin secondary structure. miRNAs are involved in post-transcriptional regulation of protein-coding genes in animals and plants by sequence complementarity within the corresponding messenger RNA. Since the discoveries of* lin-4 *and* let-7 *[1,2] in* Caenorhabditis elegans*, the number of published miRNAs in* miRBase *[3] has grown to* 8273 *entries of mature miRNAs products in primates, rodents, birds, fish, worms, flies, plants and viruses. Experimental identification of novel miRNAs is difficult because of their expression in specific conditions or cell types and only 4518 of the 8273 entries have been experimentally verified.*

Several computational methods were developed to detect new miRNAs. A first family of methods finds sequences which are homologous to known pre-miRNAs and which respect physical and geometrical characteristics of pre-miRNAs hairpin structures [4,5,6,8]. These characteristics concern the specific binding of the miRNA with his complement (miRNA) coded in the hairpin structure, secondary structural stability and hairpin structure of consensus fold between the predicted pre-miRNA and known ones. A second family of methods uses more restrictive conditions involving thermodynamically stable pre-miRNA hairpins and characteristic patterns of sequence conservation [10,11,13,14,15]. Tools have been developed to confirm pre-miRNAs based on machine-learning algorithms [7,9,17] or on the ranking of euclidian distances in a multi-dimensional space constructed from more than 30 parameters capturing the structure of pre-miRNAs [12].*

As other methods, we present a genome-wide search algorithm. It looks for homologous miRNA sequences and explores a multidimensional space, based on only 5 (physical and combinatorial) parameters. The method has been applied to a pool of phylogenetically distant genomes using a large set of already known miRNAs (all entries of miRBase). The first step of the algorithm searches for similar sequences in the genome under study by using an approximate string matching algorithm derived from [21]. Then, given a putative miRNA, it searches, using an adapted implementation of the RNAfold [16] algorithm, for putative pre-miRNAs containing it and satisfying some thresholds for the length of the sequence, for the miRNA-miRNA bonds condition and for AMFE (*AMFE$= \frac{MFE}{length} * 100$*, where MFE stands for Minimum Free Energy) as in previous approaches. It exploits two more parameters, the number of stems in the secondary structure and a MFEI threshold (*MFEI$= \frac{AMFE}{\%GC}$*) [19].*

The method is strikingly simple. Criteria defining a pre-miRNA are few but they turn out to be powerful at least as much as more sophisticated methods, like machine-learning algorithms, used to confirm pre-miRNA sequences. We applied our criteria on the data set used in [17] and [18] from Homo sapiens. We trace the Receiver Operating Characteristic (ROC) curves on these data set (varying MFEI and AMFE thresholds) and show

that our criteria discriminate pre-miRNAs from other sequences. Results are comparable to those using machine-learning, like in MiPred approach [17]. Numerically, we have 89.81% *(resp.* 93.21% *for MiPred) of specificity,* 92.0% *(resp.* 89.35%*) of sensitivity,* 90.72% *(resp.* 91.29%*) of accuracy and a Matthew's correlation coefficient (MCC) [20] of* 0.822 *(resp.* 0.826*) for parameters which optimize MCC. Based on the same intervals of optimized values, analyses were run on data set from* Arabidopsis thaliana*,* C. elegans*,* Oryza sativa *and* Rattus norvegicus *with similar performance. A similar comparison on all experimentally validated pre-miRNAs described in miRBase is in process. An exploratory analysis on eukaryotic species where no or few miRNAs are known experimentally is also running.*

In conclusion, our methodology detects new miRNAs which are similar in sequences to already known miRNAs. It demonstrates that machine-learning is not a necessary algorithmic approach for pre-miRNAs computational validation. In particular, our criteria applied to pre-miRNAs including new miRNAs is done by using only five parametrized thresholds. We obtain very satisfactory sensitivity and specificity (as shown in the comparison with machine-learning methods) for our system. The adjustment of the parameters allows us to calibrate specificity and sensitivity and this is a key feature for predictive systems (that is not present in machine learning approaches). Also, it allows us to adapt our search to pre-miRNAs which can be different from already known ones. In contrast, machine-learning can only confirm pre-miRNAs which look alike known ones, this being a limitation while exploring species with no known pre-miRNAs.

Keywords: microRNA, homologs, computational identification.

References

[1] R.C. Lee, R.L. Feinbaum, and V. Ambros, *Cell,* 75:843-854, 1993.
[2] B.J. Reinhart, F.J Slack, M. Basson, A.E. Pasquinelli, J.C. Bettinger, A.E. Rougvie, H.R. Horvitz, and G. Ruvkun, *Nature*, 403:901-906, 2000.
[3] S. Griffiths-Jones, H.K. Saini, S. van Dongen, and A.J. Enright, *Nucleic Acids Res.*, 36:D154-D158, 2008.
[4] T. Dezulian, M. Remmert, J.F. Palatnik, D. Weigel and D.H. Huson, *Bioinformatics*, 22:359-360, 2006.
[5] J. Hertel, M. Lindemeyer, K. Missal, C. Fried, A. Tanzer, C. Flamm, I.L. Hofacker, P.F. Stadler and The Students of Bioinformatics Computer Labs 2004 and 2005, *BMC Genomics*, 7:15 [epub], 2006.
[6] M. Legendre, A. Lamber and D. Gautheret, Bioinformatics, 21:841-845, 2005.
[7] A. Sewer, N. Paul, P. Landgraf, A. Aravin, S. Pfeffer, M.J. Brownstein, T. Tuschl, E. van Nimwegen and M. Zavolan, *BMC Bioinformatics*, 6:267 [epub], 2005.
[8] M.J. Weber, *FEBS J.*, 272:59-73, 2005.
[9] C. Xue, F. Li, T. He, G. Liu, Y. Li and X. Zhang, *BMC Bioinformatics*, 6:310 [epub], 2005.
[10] L.P. Lim, N.C. Lau, E.G. Weinstein, A. Abdelhakim, S. Yeka, M.W. Rhoades, C.B. Burge and P.B. Bartel, *Genes & Development*, 17:991-1008, 2003.
[11] U. Ohler, S. Yekta, L.P. Lim, D.P. Bartel and C.B. Burge, *RNA*, 10:1309-1322, 2004.
[12] Y. Xu, X. Zhou and W. Zhang, *Bioinformatics*, 24:i50-i58, 2008.
[13] X. Wang, J. Zhang, F. Li, J. Gu, X. He, T. Zhang and Y. Li, *Bioinformatics*, 21:3610-3614, 2005.
[14] J. Hertel and P.F. Stadler, *Bioinformatics*, 22:e197-e202, 2006, ISMB 2006.
[15] E.C. Lai, P. Tomancak, R.W. Williams and G.M. Rubin, *Genome Biol.*, 4:R42 [epub], 2003.
[16] I. Hofacker, W. Fontana, P. Stadler, S. Bonhoeffer, and P. Schuster, *Monatsh Chem*, 125:167–168, 1994.
[17] J. Peng, W. Haonan, W. Wenkai, M. Wei, S. Xiao, and L. Zuhong, *Nucleic Acids Res*, 2007.
[18] C. Xue, F. Li, T. He, G.P. Liu, Y. Li, and X. Zhang, *Bioinformatics*, 6:310, 2005.
[19] B.H. Zhang, X.P. Pan, S.B. Cox, G.P. Cobb, and T.A. Anderson, *Cellular and Molecular Life Sciences*, 63-2:246–254, 2006.
[20] B.W. Matthew, *Biochim. Biophys. Acta.*, 405:442-451, 1975.
[21] E.W. Myers, *Journal of the ACM*, 46(3):395-415, 1999.

Statistical modelisation of protein-ligand interaction

Stéphanie Pérot, Olivier Sperandio, Bruno Villoutreix and Anne-Claude Camproux

MTi, UMR 973 Inserm-Paris 7
35 rue Hélène Brion - 75205 Paris Cedex 13 France
firstname.name@univ-paris-diderot.fr

Abstract: *Small molecules and drugs usually interact with proteins by binding in cavities on the protein surface. Prediction methods are able to detect more or less accurately these binding-sites, but they don't describe them in term of geometrical or physico-chemical properties. However binding-sites seem to have intrinsic properties that theoretically could set them apart from other protein cavities. In this study, we try to identify geometrical and physico-chemical properties discriminating binding-sites cavities from other ones. This project could have a significant role in the process of drug discovery.*

Keywords: Drug-design, druggability, bioinformatics, chemoinformatics.

Accurate identification of protein binding-sites provides a basis for many structure-based drug design applications and protein-ligand docking algorithms. Identifying the geometrical and physico-chemical properties of ligand binding-sites plays a crucial role in virtual screening of ligands against protein structure that is widely used for the discovery of new therapeutical compounds. Given the rapidly increasing number of high resolution protein structures, it has become of great importance to develop analytical tools that identify potential binding-sites (figure 1).

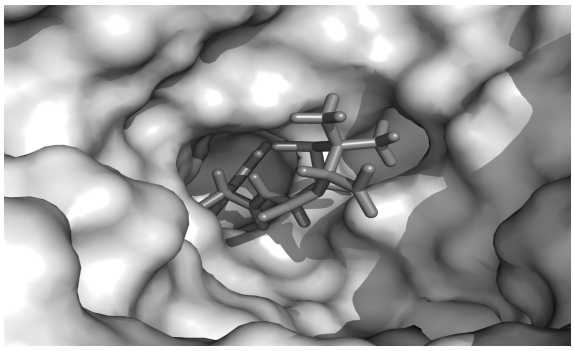


Figure 1. A ligand in its cavity.

Such prediction methods already exist. Generally, they analyse protein geometry only. Empirical studies then show that true binding-site usually coincides with the largest pockets. Among these methods we find Ligsite [1], Pass [2] or PocketPicker [3]. Other methods analyse energetic citeria and rank cavities according to the greatest energy : Q-SiteFinder [4] or PocketFinder [5]. All of these methods can predict binding-site localisation and give some description of its size and shape. PocketPicker

[3] computes a buriedness index and a rough estimate of the volume. PocketFinder [5] predict the envelope and volume of binding-sites cavities. Better knowledge of binding-sites properties should improve protein-ligand docking algorithm in speed and efficacity. So we try to determine in this study which properties, among geometrical and physico-chemical ones, are significatively different between binding-sites cavities and other protein cavities.

With this aim, we first discover two sets : a druggable set composed of drug binding-site cavities, and a non-druggable set composed of cavities that are not able to interact with drugs but are identified as potential binding-sites by prediction methods. We then compute several geometrical and physico-chemical descriptors such as volume, surface area, amino acids composition, polarity or hydrophobicity. We then apply diverse statistical methods such as support vector machine and logistic regression to find which properties contribute to discriminate cavities from binding-sites.

In agreement with An et al. [5], it was found that the volume is of great importance to discriminate binding-site cavities from other protein cavities : binding-sites are larger than other sites (figure 2, left). On the contrary, the charge show no significant difference in the two sets (figure 2, right). A predictive model combining volume, polarity and amino acids accurately predicts potential targets for drug design with an accuracy rate greater than 85% which is a quite relevant rate compared with Nayal et al. [6]. At present new descriptors such as rugosity or compacity are under consideration.

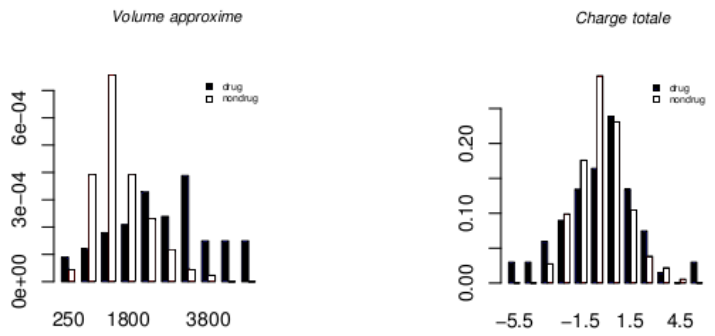


Figure 2. Example of descriptors : volume (left) and charge (right). Black is for the druggable set, white for the non-druggable set.

References

[1] M. Hendlich, F. Rippmann and G. Barnickel, LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics & Modelling*, 6:359-363,389, 1997.

[2] G. Brady and P. Stouten, Fast prediction and visualization of protein binding pockets with PASS. *Journal of Computer-Aided Molecular Design*, 4:383-401, 2000.

[3] M. Weisel and E. Proschak and G. Schneider, PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central Journal*, 1:7, 2007.

[4] A. Laurie and R. Jackson, Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, 9:1908-1916, 2005.

[5] J. An, M. Totrov and R. Abagyan, Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & Cellular Proteomics: MCP*, 6:752-761, 2005.

[6] M. Nayal and B. Honig, On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, 63:1097-0134, 2006.

Graphical development environment for bioinformatics protocol

Hervé Souiller[1,2], Sébastien Duplant[2], Yann Dantal[2], Jean-Pierre Réveilles[1]

[1] LAIC, Université d'Auvergne, IUT Clermont-Ferrand, Campus des Cézeaux, 63172 Aubière Cedex
[2] Soluscience SA, Biopôle Clermont-Limagne, 63360 Saint-Beauzire

Abstract: *The use of computing scienceis now almost obligatory part of research in* Life Sciences*, either for extraction or analysis of experimental results. In this context, several* in silico *tools are used sequentially or in parallel. The project's* **E**nv**P**rog *development platform is also born in this context. The aim was twofold:*
– *to provide a graphical interface to design a bioinformatic protocol .*
– *to provide a* compiler *allowing the transformation of the designed protocol rawing into an automaton.*
The two successive phases of developments have led to a first working version of **E**nv**P**rog.

Keywords: Bioinformatics protocols automation, software linkage.

1 Introduction

La recherche en *Sciences du Vivant* impose aujourd'hui le recourt quasi-systématique à différents outils informatiques (algorithmiques, statistiques) pour l'extraction et l'analyse des résultats d'expériences. Ces outils sont utilisés soit de manière séquentielle, soit en parallèle. L'utilisateur (chercheur, technicien,...) doit alors gérer manuellement les flux de données entre chacun des ces outils de manière à réaliser le protocole défini. C'est dans ce contexte qu'est né le projet de développement de la plateforme ***E**nv**P**rog*. L'objectif de ce projet est de fournir une interface graphique permettant de dessiner un protocole bioinformatique. La compilation de ce dessin permet la création d'un automate informatique permettant l'exécution dudit protocole.

Les développements se sont déroulés en deux phases :

le gestionnaire d'exécution : cet élément est capable de gérer le bon déroulement d'un automate de manière à mener à bien l'analyse souhaitée. Les premiers tests ont été effectués à partir de protocoles codés directement en C++.

l'interface graphique avec le *compilateur* : cette deuxième phase de développement s'est déroulée en gardant à l'esprit de fournir une interface simple permettant à un *non informaticien* de créer un programme lui-même.

2 Gestion d'exécution de *pipelines* : package *LASTec*

Afin de faciliter la conception d'un pipeline et de conserver des séquences de traitement comme une étape globale, le graphe d'exécution intègre la notion de hiérarchie où un nœud (composant d'analyse) peut être constitué de plusieurs autres nœuds. Ainsi, l'utilisateur peut *voir en profondeur* afin de concevoir son chaînage. Le notion de hiérarchie implique de décomposer la gestion de l'exécution des actions en deux entités distinctes : la séquence d'exécution et le gestionnaire d'exécution.

La séquence d'exécution est un vecteur de composants d'analyse avec une définition des dépendances entre chaque composant. Elle peut être apparentée au *design* d'un sous-modèle de traitements.

Le gestionnaire d'exécution lance et détruit les modules successivement en fonction des contraintes d'exécution. Il est capable de faire *travailler ensemble* tous les types de modules informatiques, c'est-à-dire qu'il gère les applications externes, les ressources algorithmiques internes, les interactions avec les fichiers et l'interrogation de serveurs par protocole HTTP. Par exemple, on peut lancer simultanément les modules A, B et C lors du démarrage de l'exécution du protocole. Puis, lorsque B se termine, il le détruit et lance D en lui fournissant les données issues de B.

3 L'environnement de développement graphique :

Cet environnement propose deux fonctionnalités principales :
– la programmation qui repose directement sur l'interface graphique (Figure : 1).
– l'exécution de protocoles bioinformatiques basé sur le *Gestionnaire d'exécution*.

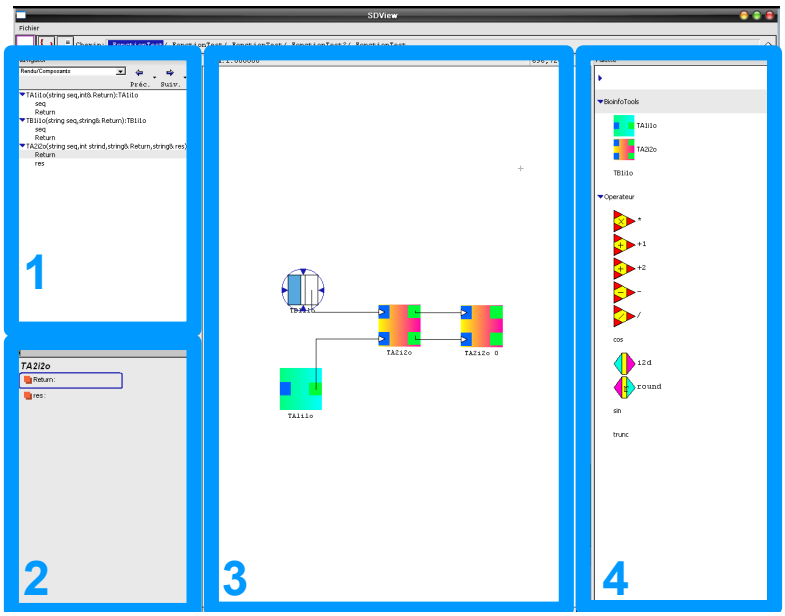


Figure 1. Interface graphique

L'interface graphique se décompose en plusieurs types d'éléments : les palettes d'outils, la zone de programmation et les éléments de gestion de données d'exécution. La zone de programmation permet de dessiner le protocole en déposant, organisant et liant des icônes représentants les outils choisis dans les palettes. Les liens créés entre les icônes symbolisent le flux des données entre les outils associés.

La compilation constiste à analyser le *dessin* fait part l'utilisateur pour générer l'automate qui sera transmis au *Gestionnaire d'exécution*. En fait, ce dessin est un graphe orienté dans lequel tous les sommets représentent des outils et les arcs, le flux des données. La compilation consiste donc à créer la séquence d'exécution à partir de ce graphe.

La gestion des données s'opère par le biais de 2 éléments graphiques : l'un pour le choix du module d'analyse, l'autre pour l'interaction (fixation/consultation avec les entrées/sorties (IO)) de l'outil choisi. En effet, chaque sommet du graphe est listé dans le premier élément et la sélection dans cette liste permet l'affichage des IO dans le second élément. L'utilisateur peut alors saisir la valeur d'une entrée et consulter les résultats après analyse.

SMALLA: a toolbox for managing libraries of smallRNA sequences

Erika Sallet[1,2] Christine Lelandais[3] Martin Crespi[3] Jérôme Gouzy[1]

[1] Laboratoire des Interactions Plantes Micro-organismes, UMR CNRS-INRA, F-31320 Auzeville
[2] PF Bioinformatique du Génopole Toulouse Midi-Pyrénées, GIS Toulouse Genopole, F-31320 Auzeville
[3] Institut des Sciences du Végétal, CNRS, F-91198 Gif-sur-Yvette Cedex and Univ. Paris VII, F-75251 Paris

Erika.Sallet@toulouse.inra.fr
Jerome.Gouzy@toulouse.inra.fr

Abstract: *SMALLA is a perl package based on the annotation platform LeARN. It provides a comprehensive suite of programs and web interfaces for analyzing the data generated by projects aiming at studying the expression of smallRNA on the basis of deep sequencing.*

Keywords: Genomics, smallRNA, miRNA, expression profiling, annotation

1 Introduction

Les technologies de séquençage actuelles rendent possible l'analyse de l'expression des molécules régulatrices que sont les smallRNA (20-25nt). Ces analyses produisent des dizaines de millions de lectures qu'il est nécessaire de filtrer pour séparer les molécules potentiellement actives des produits de dégradation. Cette étape accomplie, il devient possible de cartographier ces séquences sur le génome d'intérêt pour identifier les gènes ayant potentiellement généré les précurseurs de ces molécules et/ou pour analyser la redondance intra banque afin de comparer les profils d'expression. Parmi l'ensemble des smallRNA, les miRNA jouent un rôle majeur dans la régulation de l'expression ou de la traduction des ARN messagers de gènes codant pour des protéines. Les précurseurs des miRNA ont une structure secondaire particulière qu'il est possible de caractériser [1], de classifier et d'annoter en utilisant les séquences générées [2] pour finalement compléter l'analyse en identifiant les ARN messagers ciblés par les miRNA [3,4].

Des boites à outils bioinformatiques ont été développées récemment, par exemple `miRDeep` [5] qui permet de gérer le processus de détection et l' « `UEA sRNA toolkit` » [6] qui propose une interface web pour répondre à une grande partie des questions posées par l'analyse des smallRNA. `SMALLA` est également une boite à outil bioinformatique qui comme l'`UEA toolkit` permet de gérer une grande partie du processus d'analyse des smallRNA et plus particulièrement des miRNA. `SMALLA` est basé sur l'architecture de LeARN [7] qui est une plateforme d'annotation destinée à l'annotation des ARN non codant pour des protéines (ncRNA) à l'échelle génomique et qui propose de nombreuses fonctionnalités pour visualiser et éditer l'annotation de ncRNA. Ainsi `SMALLA`, au contraire de l' « UEA toolkit », peut d'une part être installé localement et paramétré selon l'organisme d'intérêt et les données disponibles, et d'autre part être utilisé pour visualiser et pour éditer l'annotation des précurseurs de microARN.

2 Principales fonctionnalités de la boite à outils SMALLA

2.1 Outils communs à toutes les classes de smallRNA

La boite à outils propose un ensemble de programmes permettant de nettoyer les séquences produites par les séquenceurs de type 454 ou solexa. Ce processus inclut l'identification et la suppression des adaptateurs, la suppression des lectures correspondantes à des ARN de transfert ou ribosomaux ainsi que la génération de statistiques sur la banque. Les lectures de longueurs sélectionnées (entre 20 et 25 nt par exemple) peuvent être cartographiées sur un génome d'intérêt par ncbi-blastn ou glint (Faraut et al.) puis filtrées sur la base du nombre de mesappariements. Le résultat de cette analyse produit des représentations graphiques de type « heatmap » générées par le logiciel `circos` ainsi que des fichiers GFF3 pouvant être intégrés dans une base `chado` ou visualisés directement via Gbrowse. L'interface utilisateur est un cgi-perl qui fournit un accès aux données et aux analyses en proposant les points d'entrée classiques que sont la recherche par identifiant, par séquence ou par similarité. De plus, l'interface propose un formulaire pour la recherche de molécules différentiellement exprimées entre différentes conditions dont la significativité est évaluée par la méthode proposée par Herbert *et al*. [9].

2.2 Outils spécifiques aux miRNA

Le package contient également des programmes qui vont permettre de déterminer les structures secondaires correspondant à des précurseurs de miRNA. La détection est basée sur le programme `mirfold` [1] dont les résultats sont filtrés, annotés et classifiés d'après les critères de la littérature [2]. Nous proposons également un programme qui permet une première analyse phylogénomique en comparant les séquences de mir matures à la banque de référence mirbase [8] afin d'extraire les profils des différents mir candidats. Cette analyse est complétée grâçe à un programme qui va filtrer les résultats produits par `miranda` [3] selon les critères définis par Jones-Rhoades *et al.* [4] pour prédire les ARN messagers qui pourraient être considérés comme des cibles potentielles du mir.

3 Disponibilité

SMALLA est disponible sous licence CECILL2 et intégré au package LeARN téléchargeable à partir de http://symbiose.toulouse.inra.fr/LeARN (version >= 1.5)

References

[1] Billoud, B., De Paepe, R., Baulcombe, D. and Boccara, M. (2005) *Biochimie*, **87,** 905-910.
[2] Meyers, B.C., Axtell, M.J., Bartel, B., Bartel, D.P., Baulcombe, D., Bowman, J.L., Cao, X., Carrington, J.C., Chen, X., Green, P.J. *et al.* (2008) *Plant Cell*.
[3] John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C. and Marks, D.S. (2004) *PLoS Biol*, **2,** e363.
[4] Jones-Rhoades, M.W., Bartel, D.P. and Bartel, B. (2006) *Annu Rev Plant Biol*, **57,** 19-53.
[5] Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S. and Rajewsky, N. (2008) *Nat Biotechnol*, **26,** 407-415.
[6] Moxon, S., Schwach, F., Dalmay, T., Maclean, D., Studholme, D.J. and Moulton, V. (2008) *Bioinformatics*, **24,** 2252-2253.
[7] Noirot, C., Gaspin, C., Schiex, T. and Gouzy, J. (2008) *BMC Bioinformatics*, **9,** 21.
[8] Griffiths-Jones, S., Saini, H.K., van Dongen, S. and Enright, A.J. (2008) *Nucleic Acids Res*, **36,** D154-158.
[9] Herbert, J.M., Stekel, D., Sanderson, S., Heath, V.L., Bicknell, R. (2008) *BMC Genomics,* **9**, 153.

Etude fonctionnelle d'un centre d'intéractions protéiques par une approche intégrée

2

Elodie Marchadier[1, 2], Leslie Aïchaoui-Denève[1], Rut Carballido-López[2], Philippe Noirot et Vincent Fromion[1]

[1] MIG – Mathématique, Informatique et Génome, INRA,
[2] Génétique Microbienne, INRA,
Bâtiments 233 et 440 - Domaine de Vilvert 78350 Jouy en Josas Cedex France
elodie.marchadier@jouy.inra.fr

Abstract: *Yeast-two hybird method allow to detect protein-protein interactions and allowed the construction of a protein interaction network in the model bacterium Bacillus subtilis. This network comprise a interaction center involving 9 proteins of unknown functions: a hub. It is located at the interface of essential cellular processes. We combine experimental and bioinformatic methods to explore the function of this hub. This integrative approach allow to take into account different types of data. In particular, analysis of transcriptomic data allowed us to detect expression correlations between genes of the "hub" and to adapt different statistical tools such as biclustering in order to affect genes of the "hub" to functional groups.*

Keywords: *Bacillus subtilis*, yeast two-hybrid, transcriptomic data, biclustering.

Le protéome qui est l'ensemble des protéines exprimées, est organisé en réseaux structurés d'interactions protéiques : l'interactome. Dans ces réseaux d'interactions, la plupart des protéines ont un petit nombre d'interactions alors que quelques protéines, appelées centres d'interactions ou hubs, ont un grand nombre de connexions. Notre projet se concentre sur une question biologique importante : comprendre la fonction biologique d'un hub. L'objet d'étude est un hub, découvert chez *Bacillus subtilis*, et qui se situe à l'interface de plusieurs processus cellulaires essentiels : la réplication de l'ADN [1], la division cellulaire [2], la ségrégation des chromosomes [3], la réponse au stress et la biogenèse de la paroi bactérienne [4]. Pour obtenir une vision globale de la fonction du hub, une démarche de biologie intégrative a été menée.

Après avoir réalisé une analyse du contexte génomique [5] des gènes codant pour les protéines du hub, une démarche de biologie intégrative a été amorcée en analysant des données transcriptomiques hétérogènes disponibles dans des bases de données publiques. L'analyse statistique de ces données a permis d'identifier des groupes de gènes co-régulés avec les gènes du hub. En première approche, l'analyse des corrélations entre l'expression des gènes à travers diverses conditions a été menée sur la base de l'utilisation classique de la statistique telle que la classification non supervisée. Cette première analyse, nous permet d'associer certains gènes du hub à des groupes fonctionnels, de valider et d'identifier des régulons. Elle nous permet aussi de mettre en évidence les limites d'une telle approche et la nécessité de recourir à des méthodes permettant

d'identifier les conditions dans lesquelles les gènes sont co-régulés. A cette fin, et apres une normalisation des données, nous avons utilisé des méthodes de bi-clustering, qui permettent d'identifier des groupes de gènes co-exprimés dans un ensemble significatif de conditions spécifiques. Parallèlement à cette analyse transcriptomique, de nouveaux partenaires des protéines du hub ont été identifiés et intégrés à l'analyse des corrélations. Il nous a donc été possible de combiner ces deux approches : l'étude du transcriptome et celle de l'interactome, l'une comme l'autre ont été menées de façon systématique à l'échelle du génome complet. L'intégration de ces deux types de données enrichies par une analyse phylogénétique nous permet d'éclairer le contexte fonctionnel de certains gènes de notre étude et d'émettre des hypothèses quant à la nature des interactions entre protéines du hub.

La génération et le traitement d'un tel jeu de données répond à des enjeux scientifiques majeurs, nécessitant la mobilisation des compétences, des connaissances, et des outils pour accéder à une compréhension plus globale du fonctionnement des organismes vivants. Le jeu de données constitué est utilisé pour mettre en œuvre d'autres méthodes statistiques ou informatiques. Tout cela nous permettra de disposer de méthodes permettant *in fine* d'extraire des informations de grands jeux de données en cours de production, ce qui constitue un enjeu majeur de la biologie intégrative.

Remerciements

A. Goelzer[1] et P. Bessières[1], l'équipe « Intégration fonctionnelle des processus cellulaires »[2] et E. Guedon[2], C. Hennequet-Antier (Recherches avicoles - INRA Nouzilly), E. Zalachas et P. Martin (Gel-PiCT - INRA Jouy-en-Josas), J. Aubert et S. Robin (INA PG - INRA), S. Huet et Brigitte Schaeffer (MIA - INRA Jouy-en-Josas) et B. Charnomordic (UMR-ASB - INRA Montpellier)

Références

[1] M.-F. Noirot-Gros et al. An expanded view of bacterial DNA replication. PNAS, 99(12): 8342-7, 2002.
[2] RA. Daniel et al. Multiple interactions between the transmembrane division proteins of Bacillus subtilis and the role of FtsL instability in divisome assembly. J. Bacteriol., 188(21):7396-404, 2006.
[3] E. Dervyn et al.The bacterial condensin/cohesin-like protein complex acts in DNA repair and regulation of gene expression. Mol. Microbiol., 51(6):1629-40, 2004.
[4] R. Carballido-López. Actin homolog MreBH governs cell morphogenesis by localization of the cell wall hydrolase LytE. Dev Cell., 11(3):399-409, 2006.
[5] M. Huynen et al. Current Opinion in Structural Biology., 10:366-370, 2000.

Updating the multiple alignment of composite gene families

Matthieu Barba and Bernard Labedan

Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris Sud,
Bâtiment 400, 91405 Orsay Cedex, France
matthieu.barba@igmors.u-psud.fr - bernard.labedan@igmors.u-psud.fr

1 Background

Up to now, the experimental approach to study protein families has been rather standard. Sound homologues are multiply aligned with an automatic tool. Then, this crude alignment is manually edited to reach an optimal alignment (*reference multiple alignment*) from which it becomes reasonable to reconstruct a phylogenetic tree. We call *composite protein family* an array of homologues that share the same (bio)chemical function while displaying a more or less wide assortment of biological functions. Accordingly, analyzing the tree of a composite family not only is useful in terms of protein history/species evolution but may help to cope with one of the main challenges in genomics, i.e. the step of functional annotation of newly sequenced genomes.

Each time a new homologous sequence is made available, the whole process has to be repeated in order to get an updated tree. With the deluge of new sequences of the complete genome of nearly any organism (around four new genomes per week), it becomes more and more demanding to add manually each new homologue to the tree of a composite family. To cope with this challenge, we designed a nearly automatic approach that starts from an initial set of experimentally characterized homologous sequences and adds in a stepwise process more and more sound homologues.

2 Continuously updating a reference multiple sequence alignment

2.1. Producing a seed alignment

Our initial set of amino acid sequences must be reliable at the level of their functional annotation. Accordingly, we collected in the Swiss-Prot database [1] protein sequences having an existence of type 1 (proteins that have been experimentally studied). These sequences are organized in a relational database and multiply aligned [2]. This multiple sequence alignment (MSA) is further manually edited to be used as a seed alignment. In particular, the automatic introduction of indels is ascertained by using information from the known tertiary structures of the sequences under study, if available. Such certification is becoming routine with the rapid increase of the number of new 3D structures that are added to the PDB. Alternatively, if enough amino acid sequences of crystallized proteins are available, we align them directly with the program 3D-Coffee [3] that has been benchmarked as optimal when sequence identity between target and template falls below 50% [4].

2.2. From the seed alignment to the reference alignment

This seed alignment is further enlarged with homologues that present a reliable level of identity with at least one of the experimentally studied proteins using the following steps:

- Each sequence of the seed alignment was used as a query in a search (Blastp program used with the Blastall feature) of partners in public databases that share at least 30% identity with a pairing extending at least 67% of the length of the shorter matching protein.
- Close homologues (sharing at least 70% identity with a pairing extending at least 67% of the length of the shorter matching protein) are further clustered using the Cd-hit program [5].
- For each cluster, a MSA is built [2] and a HMM profile is computed (*HMM_cluster*). In parallel, another HMM profile is computed for the set of sequences present in the seed alignment that are the closest ones to this cluster (*HMM_seed*). Then, the two profiles

HMM_cluster and *HMM_seed* are aligned using the HHalign program [6].

- A stepwise approach allows adding progressively each aligned cluster to the seed alignment. Starting with highly identical sequences and repeating the whole process by progressively lowering the identity threshold down from 60 to 55, 50, 45, 40, and 30 makes the update of the MSA more efficient and safer. In particular, this help to limit the addition to the eventually obtained reference alignment of new indels that are not compatible with information based on the seed alignment.

2.3. Updating the reference alignment and the phylogenetic tree

We can further update this optimal MSA by adding, as soon as they are published, any new homologous sequence by reiterating the approach described above. Accordingly, one can reconstruct the most recent phylogenetic tree [7], allowing to refine the study of protein history and species evolution and to improve by monophyly the functional annotation of the added homologues.

3 A test case

To illustrate the efficiency of our tool, we applied it to a badly defined composite family. We chose dihydroorotases (third step of pyrimidine biosynthesis) as a test case because (i) this enzyme family is rather complex with two types separated in several classes [8], (ii) we previously showed that some of these types interact with aspartate carbamoyltransferases (second step of pyrimidine biosynthesis) at the level of their quaternary structures [9], (iii) these dihydroorotases belong to the superfamily of amidohydrolases [10]. Using our approach, we obtained a tree that allows to (re)annotate many of these amidohydrolases and to precise the type and class of each dihydroorotase. Over the past two years, the tree kept its global topology at each update, allowing many new unknown homologues to be correctly annotated or reannotated. We will demonstrate the efficiency of our approach by using specific statistical tools [11,12] to compare the tree made with our tool and the standard tree obtained when aligning in one step the same set of sequences.

References

[1]The UniProt Consortium. The Universal Protein Resource (UniProt). Nucleic Acids Res. 36:D190-D195, 2008

[2]Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5:113, 2004.

[3] O'Sullivan,O. Suhre K, Abergel C, Higgins DG, Notredame C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J. Mol. Biol., 340:385–395, 2004

[4] Dalton JA, Jackson RM. An evaluation of automated homology modelling methods at low target template sequence similarity. Bioinformatics 23:1901-1908, 2007

[5] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences Bioinformatics 22:1658-1659, 2006

[6] Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951-960, 2005

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

[8] Fields, C., Brichta, D., Shepherdson, M., Farinha, M. and O'Donovan, G.A. Phylogenetic analysis and classification of dihydroorotases: a complex history for a complex enzyme. Paths to Pyrimidines. An International Newsletter 7:49-63, 1999

[9] Labedan B., Xu Y., Naumoff D. G. & Glansdorff N. Using quaternary structures to assess the evolutionary history of proteins: the case of the aspartate carbamoyltransferase. *Mol. Biol. Evol*. 21:364-73, 2004.

[10] Holm, L. and Sander, C. Unification of a broad set of amidohydrolases related to urease. Proteins 28:72-82, 1997

[11] Cantarel BL, Morrison HG, Pearson W. Exploring the relationship between sequence similarity and accurate phylogenetic trees. Mol Biol Evol. 23:2090-2100, 2006

[12] Soria-Carrasco V, Castresana J. Estimation of phylogenetic inconsistencies in the three domains of life. Mol Biol Evol. 25:2319-2329, 2008

Paysage d'énergie et structures localement optimales d'un ARN

Azadeh Saffarian[1,2], Mathieu Giraud[1,2], Hélène Touzet[1,2]

[1] LIFL, UMR CNRS 8022, Université Lille 1, France
[2] INRIA Lille Nord-Europe, France

Abstract: *To understand the functional role of an RNA in the cell, it is useful to know its structure or at least an approximation of its structure such as the secondary structure. In this perspective, the energy landscape gives a useful insight into all potential conformations of the molecule. Here we propose to study the energy landscape of a given RNA sequence by considering locally optimal structures. Locally optimal structures are thermodynamically stable structures maximal for inclusion: they cannot be extended without producing a conflict. We propose a new algorithm,* `Regliss`*, that computes all locally optimal structures for an RNA sequence. The algorithm is implemented and runs on a publicly accessible web server.*

Keywords: RNA, energy landscape, secondary structure, locally optimal structures

1 Structures localement optimales

Pour découvrir et comprendre la fonction d'une molécule d'ARN, comme un ARN non-codant, il est utile de connaître sa structure, ou au moins une description de sa structure comme la structure secondaire. Dans ce contexte, le *paysage énergétique* d'un ARN fournit une information riche: c'est l'ensemble des conformations possibles et leur énergie.

Peu d'outils existent pour étudier ce paysage énergétique. `Mfold` [8,4] prédit la structure optimale d'un ARN, ainsi que certaines structures sous-optimales. Cet ensemble ne couvre toutefois pas l'intégralité des structures sous-optimales. Pour chaque paire de bases, `Mfold` génère en effet la meilleure conformation contenant cette paire de bases. Certaines structures contenant deux appariements non optimaux ne sont ainsi pas obtenues. Alternativement, `RNAsubopt` [7] produit *toutes* les structures secondaires d'une séquence d'ARN. Vu le nombre exponentiel de ce type de structures, les informations intéressantes pour obtenir un paysage énergétique sont alors noyées dans un grand nombre de structures inutiles. À cet effet, `RNAshapes` [6] est un post-traitement de `RNAsubopt` qui réduit ce grand nombre de structures à quelques structures globalement différentes.

On voit que malgré ces méthodes existantes, décrire d'une manière concise et exacte l'ensemble des conformations pertinentes d'une molécule d'ARN reste un problème non résolu. Comment choisir des représentants pertinents parmi toutes les conformations possibles ? Les *structures localement optimales* sont des structures secondaires maximales par l'inclusion: dans ces structures, tout nouvel appariement entre deux bases conduit à la création d'un pseudo-nœud ou d'un triplet. Ce concept a déja été étudié par Clote [1], qui a proposé un algorithme de dénombrement de telles structures. À côté de cette définition topologique, a été introduite une définition énergétique [3]: les structures sont les minimums locaux du paysage énergétique sur le modèle de Turner, et cet ensemble contient la structure optimale.

2 `Regliss`: algorithme et implémentation

Nous proposons une approche qui s'inspire de la définition topologique, tout en prenant en compte les informations énergétiques, à travers la formation d'hélices thermodynamiquement stables. Pour cela, nous considérons des structures localement optimales construites *à partir d'un ensemble de tiges potentielles* (fournies par exemple par `unafold` [4], `mc-fold` [5] ou représentant des contraintes connues par l'utilisateur). Les relations entre deux tiges potentielles sont: juxtaposition, croisement, inclusion ou conflit. Une structure localement optimale est ici une structure contenant le maximum de tiges compatibles: on ne peut plus ajouter de tiges sans créer un conflit.

Nous avons conçu un algorithme, appelé `Regliss` (pour *RNA energy landscape and secondary structures*), pour ce problème. L'idée de l'algorithme est que l'ensemble des structures localement optimales peut être généré à partir de l'ensemble des structures *maximales par juxtaposition*, sans relation d'imbrication. Pour produire ce type de structures, l'efficacité du calcul est assurée par deux propriétés. En premier lieu, l'algorithme procède par programmation dynamique, construisant l'ensemble des structures maximales pour la juxtaposition de manière incrémentale, à partir des résultats sur les sous-séquences. Le deuxième point est qu'il dispose d'un mécanisme de filtrage qui évite de reproduire les structures déjà incluses dans une autre structure précédemment trouvée.

Il est à noter que la définition d'optimalité locale est principalement topologique, et permet de s'affranchir d'un modèle énergétique sophistiqué. Nous proposons toutefois, en option, de trier et de sélectionner les structures localement optimales suivant l'énergie libre fournie par le modèle thermodynamique standard avec `RNAeval` [2]. `Regliss` a été implementé en C. Un serveur web est disponible à l'adresse `http://bioinfo.lifl.fr/RNA/regliss`, qui offre également des outils de visualisation des structures localement optimales.

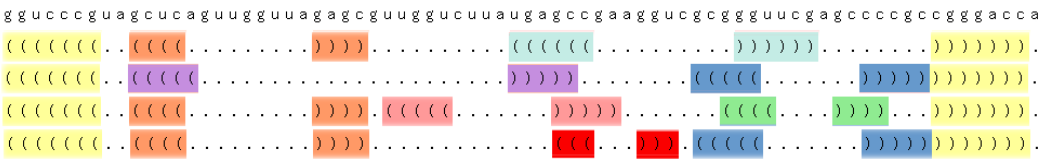


Figure 1. Exemple de résultat de `Regliss`.

References

[1] P. Clote. An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov-Jacobson energy model. *J. Computational Biology*, 1:83–101, 2005.

[2] I. L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte F. Chemie*, 125:167–188, 1994.

[3] W. A. Lorenz and P. Clote. Calculating local optima in the turner energy model for rna secondary structure. *ISMB 2008, Poster session Q34*, 125, 2008.

[4] N. R. Markham and M. Zuker. Unafold: software for nucleic acid folding and hybriziation. *Bioinformatics*, 2, 2008.

[5] M. Parisien and F. Major. The mc-fold and mc-sym pipeline infers rna structure from sequence data. *Nature*, 425:51–55, 2008.

[6] P. Steffen, B. Voss, M. Rehmsmeier, J. Reeder, and R. Giegerich. Rnashapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.

[7] S. Wuchty, W. Fontana, IL. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–165, 1999.

[8] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, 1989.

RNAspace: non-coding RNA annotation web platform

Philippe Bardou[1], Marie-Josée Cros[2], Christine Gaspin[2,3], Daniel Gautheret[4],
Jean-Marc Larré[3], Benjamin Grenier-Boley[5], Jérôme Mariette[3], Antoine de Monte[5], Hélène Touzet[5]

[1] INRA, SIGENAE, UMR 444, F-31320 Castanet,
[2] INRA, Unité de Biométrie et Intelligence Artificielle, UR 875, F-31320 Castanet,
[3] INRA, Plateforme bioinformatique, F-31320 Castanet,
[4] IGM UMR 8621 CNRS-U Paris sud,
[5] LIFL, UMR CNRS 8022 Université Lille 1 and INRIA LNE
contact@rnaspace.org

Abstract: *RNAspace is a web platform for non-coding RNA (ncRNA) identification in genomic sequence. The platform offers an integrated environment for running a variety of ncRNA gene finders, exploring predictions with dedicated tools for comparison, visualization and edition of candidate ncRNAs and exporting results in various formats. The platform is developed both as a web site (with limitations on analyzed sequence size and execution time), and for local installation with user authentication.*

Keywords: non coding RNA, annotation, ncRNA gene finder.

The availability of complete genome sequences and the development of high throughput technologies have led to the accumulation of raw biological data at an unprecedented scale. Whereas structural and functional protein annotation is now considered as a task which is relatively well solved, ncRNA genes are not (or at a weak level) integrated in these environments. This fact can be explained by a few reasons which are respectively a recent interest for ncRNA, the absence of general ncRNA prediction methods and the difficulty to analyze these molecules with regard to their sequence and structure conservation. The latter task generally requires an expertise level not widespread and the need to use analysis and edition tools more sophisticated than pure similarity search. The increasing number of ncRNA discovered and the lack of user friendly tools for finding and annotating them, have made necessary to propose to biologists an *in silico* environment allowing structural and functional annotations of these molecules.
For this purpose, a web platform called RNAspace is being developed as a collaborative and open software allowing to:

- run a variety of ncRNA gene finders in an integrated environment,
- explore computed results with dedicated tools for comparison, visualization, alignment and edition of putative ncRNAs
- and export them in various formats (FASTA, GFF, RNAML).

Gene finders are organized into three categories containing respectively :

1. known RNA based gene finders including (i) sequence homology search tools (BLAST [1], YASS [2]) on ncRNA databases (Rfam [3], fRNAdb [4]), (ii) general purpose RNA motif search tools (darn! [5], Erpin [6]), (iii) specialized search tools (RNAmmer for ribosomal RNAs [7], tRNAscan_SE for transfer RNAs [8])
2. comparative analysis gene finders (an *ad hoc* pipeline has been implemented based on BLAST or YASS for similarities search and caRNAc [9] or RNAz [10] for consensus structure inference),
3. an *ab initio* gene finder based on detection of atypical GC% regions.

Once the execution of selected gene finders is achieved, an overview of all putative ncRNAs found on the genomic sequence is given. Their main characteristics are displayed in a list that can be dynamically explored by sorting and filtering its content. For each putative ncRNA or a selection of them, more details are computed on line (*e.g.,* compute and visualize a secondary structure, align a selection of predictions and visualize the alignment, save and store an alignment...). It is also possible to edit and to delete any putative ncRNA.

The platform is developed to be both available through a web site www.rnaspace.org (with limitations on analyzed sequence size and execution time) and local installations with user authentication. RNAspace is a collaborative platform, that is intended to be in constant development. In the near future, we plan to incorporate supplementary prediction approaches, to provide advanced tools to eliminate redundant results, to add visualization features using a genome browser, to include information on the genomic context...

Acknowledgements

This work is founded by RNG (Réseau National des Génopoles), french network of Genomic Centers.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. J Mol Biol 215 (3): 403–410, 1990.

[2] L. Noe, G. Kucherov, YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Research, 33(2), 2005:

[3] P.P. Gardner, J. Daub, J.G. Tate, E.P. Nawrocki, D.L. Kolbe, S. Lindgreen, A.C. Wilkinson, R.D. Finn, S. Griffiths-Jones, S.R. Eddy and A. Bateman, Rfam: updates to the RNA families database. Nucleic Acids Research, 37(Database Issue), 2009.

[4] T. Kin, K. Yamada, G. Terai, H. Okida, Y. Yoshinari, Y. Ono, A. Kojima, Y. Kimura, T. Komori, K. Asai, fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. Nucleic Acids Research, 35(Database Issue), 2007.

[5] M. Zytnicki, C. Gaspin, T. Schiex, DARN! A Weighted Constraint Solver for RNA Motif Localization. Constraints, Vol. 13, 2008.

[6] D. Gautheret, A. Lambert, Direct RNA Motif Definition and Identification from Multiple Sequence Alignments using Secondary Structure Profiles. J Mol Biol. 313:1003-11, 2001.

[7] K. Lagesen, P. Hallin, E.A. Rødland, H.-H. Stærfeldt, T. Rognes, D.W. Ussery, RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research, 35(9), 2003.

[8] T.M. Lowe, S.R. Eddy, tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research, 25(5), 1997.

[9] H. Touzet, O. Perriquet, CARNAC: folding families of non coding RNAs. Nucleic Acids Research, 142(Web Server Issue), 2004.

[10] S. Washietl, I.L. Hofacker, P.F. Stadler, Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. U.S.A. 102, 2454-2459, Feb. 2005.

A new portal on INRA URGI bioinformatic platform, to bridge genetics and genomics plant data with 2 new tools, a quick search tool and an advanced search tool to mine the data

Delphine Steinbach[1], Erik Kimmel[1], Aminah-Olivia Keliet[1], Michael Alaux[1], Nacer Mohellibi[1], Daphné Verdelet[1], Joelle Amselem[1], Sophie Durand[1], Cyril Pommier[1], Isabelle Luyten[1], Sébastien Reboux[1], Hadi Quesneville[1]

[1]URGI (Unité de Recherche Génomique-Info, INRA,
Centre INRA de Versailles, bâtiment 18, RD 10, 78000, Versailles, France
urgi-contact@versailles.inra.fr

Abstract: *URGI (Unité de Recherche Génomique-Info) is an INRA bioinformatics unit dedicated to plants and pest genomics. Created in 2002, one of its mission is to develop and host a genomic and genetic information system called GnpIS, for INRA plants of agronomical interest and their bioagressors. It hosts a bioinformatics platform which belongs to the ReNaBi network and has a national interorganism label (RIO/IBISA 2007). The URGI maintains an efficient computing environment and offers services covering database conception, software engineering, and bioinformatics. Since 2008, a focus is done on doing an interoperability between both the tools (a set of Oracle, PostGreSql, MySql databases and their interfaces in Java,Perl) and the data located in all these databases (for example on Wheat and Gravevine data). New developments and 2 new tools were developped since march 2008 to search through all these data like a "google search" via indexes ("**Quick search tool**") or via dedicated marts ("**Advanced search tool**") We will present here these last developments, already available since January 2009 on our public web site at this url:http://urgi.versailles.inra.fr/gnpis. In summary ,it is the first version of a new portal to bridge plant genetics and genomics data, which rely on a set of databases (GnpIS) and 2 new tools to query through all these databases transparently, one tool ("a quick search tool") based on Lucene (a high-performance, full-featured text search engine library, http://lucene.apache.org,), the other one ("an advanced search tool") based on Biomart (a query-oriented data management system http://www.biomart.org)*

Keywords: Genomics, databases, information system, bioinformatics platform, index, lucene, mart

1 Introduction

INRA Unité de Recherches Génomique-Info (URGI) is a bioinformatics research INRA unit created in 2002 (previously known as Genoplante-info). The unit has a research team which works on genome structure, dynamics and evolution by focusing on repeat sequences analysis (REPET pipeline). The unit hosts a bioinformatic platform, labelled at national level: RIO/IBISA 2007 and member of the French National Network of Bioinformatic Platforms (ReNaBi). The main platform mission is to maintain a repository called GnpIS for plant and pest genomic and genetic data and to offer tools, web interfaces, pipelines and support to biologists and bioinformaticians to mine and extract valuable information in a single repository to be able to navigate, analyze, compare and export data. The platform is since 2000, the official repository for Genoplante projects data. It is also now the repository for Wheat and Grape data at INRA level but also at international level for Grape sequence annotation data. It is in increase in terms of size and projects number as coordinator or partners. It maintains a private site for these partners (https://gpi.versailles.inra.fr, ask urgi-contact@versailles.inra.fr for an account) and for all a public Web site: (http://urgi.versailles.inra.fr).

The results presented in this poster are focused on the last developments done since 2008, in the frame of an ANR GnpInteGr project, whose aim is to build a new portal to be able i) to query transparently and rapidly (as "a google search") through genomics and genetics data located in all the databases (GnpIS) of the URGI platform information system and also to be able ii) to build complex queries over dedicated marts, made according to biologist needs in fields like genetic and physical mapping, polymorphism and genome annotations.

2 GnpIS, information system :

The URGI information system called GnpIS, is a web based system composed of a several applications (in Java and Perl) built above a centralized relational database that includes schemas dedicated to sequence data (EST, contigs) in GnpSeq database module, genomic annotation data in GnpGenome, genetic mapping data (markers, maps, QTLs) in GnpMap, expression data in GnpArray, proteomic data in GnpProt, SNP data in GnpSNP but also genetic resources data and phenotypes in Siregal, Ephesis applications. Data are submitted by the laboratories through an automatic Web submission tool which allows the checking and the data bulk loading. Web interfaces allow the biologists to query and visualize the data and navigate through them.

The 2008-2009 ongoing developments are the creation of an **interoperability between the genomic and genetic databases modules to allow either a quick simple query like "a google search", either integrated queries based on dedicated datamarts, involving all kind of data together and providing both as results, a list of items allowing to go deeper into details in the data via the existing Gnp* interfaces**. 2 technologies are used, Biomart and Hibernate/Lucene technology, JAVA J2EE technology. We will present in this poster, the first version of this new portal, on line since December 2008 for its first version on **http://urgi.versailles.inra.fr/gnpis**. This interface is improved and released regularly (following Agile development methodology) to add new functionalities according to users feedbacks and user needs. Some demo and tutorial (electronic tutorial via animation) are available. New important data are also in progress to be loaded into the information system, to provide more useful links to biologists in their research fields to identify for example, genes responsible for their traits of interest. **3 'interoperable' pilots data sets ' are in the way to be released on the public site: a poplar, a grapevine and a wheat sets.** A publication is in preparation.

3 Citations

Acknowledgements

This work is supported by Genoplante from 2000 to 2005, ANR Genoplante 2007 (GnpInteGr, GnpAnnot projects) and INRA (GAP and SPE departments). We thank all the contributors to the development of the URGI bioinformatics platform and information system: data productors (biologists), data submitters (biologists, bioinformaticians), people involved in working groups and developers. Especially, we thank all the URGI team and our collegues, Sandra Derozier who begins the work on biomart, Philippe Leroy, Catherine Ravel, Etienne Paux, Frédéric Choulet and Catherine Feuillet for wheat data genomics, Anne-Françoise Adam-Blondon, Nathalie Choisne for grapevine genomics, Dominique Brunel and Fabienne Granier for SNP, Alain Charcosset, Johann Joets for genetic mapping and maize data, Marc-Henri Lebrun and Thierry Rousselle for fungi genomics, Patricia Faivre Rampant, Isabelle Bourgait, Véronique Jorge for poplar data, Christophe Plomion for tree genomics, Stéphanie Sidibe-Bocs, Manuel Ruiz from CIRAD and rice data and Hélène Lucas, Catherine Christophe from INRA strategic support and fundings.

References: See. http://urgi.versailles.inra.fr/about/publications

The URGI plants and bio-agressors genomic annotation system

Baptiste Brault[1,2], Michael Alaux[1], Fabrice Legeai[3], Sébastien Reboux[1], Isabelle Luyten[1], Stéphanie Sidibe-Bocs[4], Delphine Steinbach[1], Hadi Quesneville[1], Joelle Amselem[1,2]

[1] INRA Unité de recherche en Génomique-Info, Route de St Cyr, 78026 Versailles Cedex France

[2] INRA Biologie et gestion des Risques en agriculture – Champignons pathogènes des plantes, Route de St-Cyr, 78026 Versailles Cedex France

[3] INRA Rennes Biologie des Organismes et des Populations appliquées à la protection des plantes 1099 Rennes AgroCampus

[4] UMR DAP - Cirad TA A-96 / 03 (Bât. 3, Bur. 12) Av Agropolis 34398 Montpellier Cedex 5

baptiste.brault@inra.versailles.fr

Abstract: *The URGI genomic annotation platform, developped in the framework of the GnpAnnot project, relies on well known GMOD tools (http://gmod.org): Apollo, Chado and GBrowse. Apollo is the graphical interface for visualization and annotation edition allowing curators to edit their genes according to evidences (transcript and protein similarity, comparative genomics). Manual annotations (gene curation validated/in progress) are saved in a dedicated Chado database and shared at the same time with other community annotation members. Validated genes/pseudogenes are then commited in a second Chado database accessible by Gbrowse. We will present here this "roundtrip" annotation system.*

Keywords: genomic annotation platform, structural annotation, functional annotation, manual curation, database, GMOD, Apollo, Chado, GBrowse, URGI, GnpAnnot.

1 Introduction

The INRA URGI (Unité de Recherche en Génomique-Info) is a bioinformatic team whose main mission is to develop and maintain an information system for plant and bio-agressors genomes through the development of national or international collaborative projects involving biologists and bioinformaticians. It maintains a public Web site: http://urgi.versailles.inra.fr
URGI is involved in the GnpAnnot project to set up a structural and functional manual annotation system for eukaryotic genomes.

2 Roundtrip Apollo – Chado – GBrowse

In the framework of the GnpAnnot project, we set up a manual genomic annotation platform relying on the international open source project Generic Model Organism Database (GMOD, http://www.gmod.org). The main goal is the development of the data flow management called "the roundtrip" between Chado database, Apollo genome annotation editor and GBrowse genome browser. Apollo is directly connected to Chado using the "pure JDBC" direct communication protocol. Moreover, we are also involved in the improvement of Apollo software.

Researchers create/curate genes using Apollo genome editor on their personal computer. Once genomic annotations are validated (genes, pseudogenes), a pipeline extracts data from Chado4Apollo database to gff3 files, then updates Chado4Gbrowse database interfaced by GBrowse available at: http://urgi.versailles.inra.fr/gbrowse.

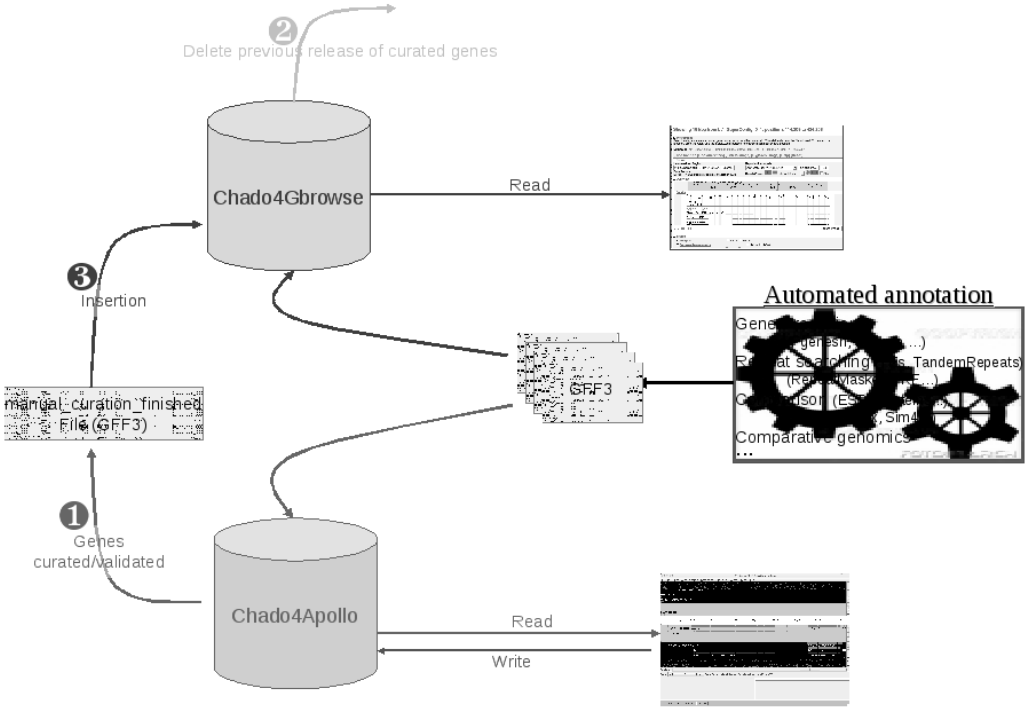


Figure 1. Roundtrip Apollo <-> Chado <-> Gbrowse

This figure shows the different steps of the roundtrip. Results obtained from automated annotation are parsed in GFF3 format. This GFF3 files are used to populate both Chado4Gbrowse and Chado4Apollo databases (Chado model). Gene manual curation is done by scientists using Apollo interface. Chado4Gbrowse is periodically updated with validated genes extracted from Chado4Apollo.

3 Application

Manual curation roundtrip system, using Apollo, has been set up for two genome communities annotators: the International Aphid Genome Consortium and the International *Botrytis-Sclerotinia* genome project consortium. We are currently setting up new instances of the system in the framework of the Franco-Australian *Leptosphaeria maculans* genome project, the Franco-Italian Tuber genome project and the International Vitis genome consortium.

4 Perpectives

In the future, we will improve the manual curation platform and develop a fonctional annotation curation system. Try to make easier the "roundtrip" with only one database for GBrowse and Apollo.

Acknowledgements

Thanks to URGI team, current and former members. Especially Cyril Pommier and Olivier Arnaiz. Thanks to GnpAnnot projects partners and GMOD consortium.

sHSPprotseqDB : a database for the analysis of small Heat Shock Proteins

Mathieu Almeida [1], Pierre Poulain [1], Catherine Etchebest [1], Delphine Flatters [1]

[1] INSERM U665, INSERM et Université Paris Diderot-Paris 7,
Equipe Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB),
INTS, 6 rue Alexandre Cabanel, 75015 Paris, France
mathieu.almeida77@gmail.com
pierre.poulain@univ-paris-diderot.fr
catherine.etchebest@univ-paris-diderot.fr
delphine.flatters@univ-paris-diderot.fr

Abstract: *The small Heat Shock Proteins (sHSP) belong to the chaperone family. Until now, only three structures of sHSP are available in the Protein Data Bank (PDB). The database sHSPprotseqDB aims at collecting the large amount of protein sequences of sHSP reported in Uniprot in order to predict their structural properties. The sHSP are ubiquitous proteins that can be classified into distinct, well-characterized groups. Combining sequence alignments, secondary structure predictions and available data reported in the literature, the three regions, Nter, ACD and Cter, are delimited for each protein. Preliminary statistical analyses reveal that it is more relevant to study separately the three regions than the whole sequence. Our first results show that these regions display specific properties in their length, amino acid composition or main secondary predicted structure. The database is intended to facilitate further correlation analysis on the sHSP. The originality of this database is to include sequence features and structural properties in order to help the biologist to design experiments, for instance mutagenesis on appropriate locations, and to gain some structural information.*

Keywords: small Heat Shock Proteins (sHSP), chaperone, database, structure prediction.

The small Heat Shock Proteins are chaperone-like proteins. Under a cellular stress, they are able to protect unfolded proteins or non native proteins against aggregation [1]. They are thus able to interact with a large variety of substrates, however the mechanisms of this chaperone-like function are still poorly described. At structural level, they form assemblies of variable size but the building subunit of the assembly has a very conserved fold within the family. This subunit is composed of three regions, namely the Nter, the Alpha Crystallin Domain (ACD) and the Cter region [2]. Whereas only 3 structures are available in the Protein Data Bank (PDB), more than 2000 protein sequences are reported in Uniprot.

We aim to collect and explore the protein sequences in order to predict structural properties in the sHSP family. In this work, we present the first database, sHSPprotseqDB, dedicated to the small Heat Shock Protein family. This database (DB) was built in several steps : (i) determination of

criteria to extract the protein sequences from Uniprot, (ii) classification of the sequences in groups as described in the literature, (iii) extraction of properties on the whole sequences or on the different regions (amino acid composition, length, secondary structure predictions ...), (iv) implementation of queries in the DB for correlation analysis.

A preliminary dataset was built containing more than 2000 protein sequences collected in Uniprot. These sequences were selected depending of their protein attributes (sequence length, protein existence level, ...) as defined in Uniprot. Different informations are also extracted directly from Uniprot annotation (accession number, organism, protein attributes, function, subcellular location ...). According to the literature, the sequences are classified into groups such as fungi, plants, animals, archaea, or bacteria [3]. Structural properties and sequence analysis are computed. All these results are combined with the three sHSP known structures to delimitate the three regions, Nter, ACD and Cter, in each sequence. Our first results show that it is relevant to study the different regions separately instead of the whole sequence. In accordance with the literature, the ACD region is the most homogeneous region with about 88 residues in length. This region is described as the signature of this protein family. In contrast, the Nter and Cter have much more variable lengths ranging from few to hundreds of residues. In term of amino acid composition, we can point out differences comparing the regions that are not detectable in the whole sequence. Finally, each region has a propensity to fold in a particular secondary structure: the structure prediction in ACD region is mainly beta strands whereas in the Nter, it is preferentially helices. At this stage of the work, we are exploring correlation analysis between groups and/or between regions, extracting conserved motifs and predicting other structural properties from sequences. All these data will be combined to give some clues from sequence to structure.

References

[1] M. Haslbeck, T. Franzmann, D. Weinfurtner, J. Buchner, Some like it hot: the structure and function of small heat-shock proteins, Nat. Struct. Mol. Biol. 12 : 842-6, 2005.

[2] D. Flatters, S. Fournier, P. Vicart and C. Etchebest, In Silico approaches to study the associative properties of the small Heat Shock Proteins (sHSPs), Abstract P-550, Proceedings of IUPAB/EBSA International Biophysics Congress in Eur. Biophys. J. , 34(6):721,2005.

[3] X. Fu, Z. Chang, Identification of a Highly Conserved Pro Gly Doublet in Non animal Small Heat Shock Proteins and Characterization of Its Structural and Functional Roles in Mycobacterium tuberculosis Hsp16.3, Biochemistry (Moscow), 71 : 583-590, 2006.

Base de données Génolevures : génomique comparative des Hemiascomycetes

Tiphaine Martin [1], David James Sherman [12], Macha Nikolski [1], Jean-Luc souciet [3], Pascal Durrens [1]
pour le Consortium Génolevures

[1] LaBRI,Laboratoire Bordelais de Recherche en Informatique,UMR CNRS 5800,
351 cours de la Libération 33405 Talence cedex France
[martin,macha,durrens]@labri.fr

[2] INRIA, Institut National de Recherche en Informatique, Centre de Recherche Bordeaux Sud-Ouest,
351 cours de la Libération 33405 Talence cedex France
david@labri.fr

[3] Université Louis Pasteur, CNRS, UMR 7156, GDR 2354, Institut de Botanique,
28 rue Goethe 67000 Strasbourg France
jean-luc.souciet@gem.u-strasbg.fr

Abstract: *The Génolevures online database (http://cbi.labri.fr/Genolevures/ and http://genolevures.org/) provides exploratory tools and curated data sets relative to nine complete and seven partial genome sequences. They were determined and manually annotated by the Génolevures Consortium, to facilitate comparative genomic studies of Hemiascomycete yeasts. The 2008 update of the Génolevures database provides four new genomes in complete (subtelomere to subtelomere) chromosome sequences, 50 000 protein-coding and tRNA genes, and in silico analyses for each gene element. A key element is a novel classification of conserved multi-species protein families and their use in detecting synteny, gene fusions and other aspects of genome remodeling in evolution. Our purpose is to release high-quality curated data from complete genomes, with a focus on the relations between genes, genomes and proteins*.

Keywords: comparative Genomics, databases, protein families, syntenies, yeast.

1 Introduction

Depuis 1999, Le consortium Génolevures explore l'évolution génomique des eucaryotes à travers une comparaison à large échelle de génomes de levures annotés manuellement. La base de données publique Génolevures évolue à chaque nouvelle version significative telle que : en 2004 avec 13 génomes partiels [1,2], en 2006 avec 4 génomes complets [3,4] et en 2008 avec 4 nouveaux génomes complets [5].

2 Les données et résultats des études

Le consortium Génolevures séquence, annote, et analyse des génomes complets venant de la branche des Hemiascomycetes et réalise sur ses données et celles provenant de génomes extérieurs, de nombreuses études de comparaison *in silico* et expérimentales.
A partir de ces comparaisons, nous produisons des classifications de gènes, de protéines et de séquences pour faciliter les études sur l'évolution moléculaire. Pour la conservation des gènes, les gènes spécifiques, nous fournissons des familles de protéines et un outil d'étude des motifs

d'épissage, Génosplicing. Par exemple, les 48 889 protéines dans les protéomes prédits des 9 génomes complets sont classées dans 7927 familles, dont 4369 sont commun à au moins 2 espèces. Pour la conservation des fonctions, nous donnons accès aux données de YETI, une classification des protéines de transport membranaires, les liens vers d'autres bases de données, aux familles de protéines, ou encore à une étude de reconstruction de voies métaboliques. Enfin pour le remodelage des génomes, nous mettons à disposition les résultats de travaux sur les blocs de synténies et les fusions de gènes.

3 L'exploration et la structuration du site

La base de données Génolevures est conçue pour fournir des outils aidant à comprendre les mécanismes d'évolution moléculaire des eucaryotes. Ainsi, les questions clés dans le cas de Génolevures sont :1) Quels gènes existent, quels sont les orthologues pour mon gène favori ou quels sont les membres d'une classe fonctionnelle (mot clé, alignement, homologie)? 2) Qu'est-ce qui est connu sur un élément chromosomique donné (élément chromosomique)? 3)Quels types de relation existent dans une famille de protéine (familles de protéines)? 4) Comment sont organisés les génomes individuellement (cartes, genome browser) et entre eux (synténie)? 5) Comment sont classifiés les gènes et les protéines (familles de protéines, fusions, tandems, intron, YETI)?. Pour aider les chercheurs à répondre à ces questions, nous fournissons un outil de recherche d'éléments via une interrogation par son nom, son annotation, son appartenance à une famille. La recherche d'élément peut aussi se faire via son positionnement dans le génome, dans ce cas-là, nous donnons accès à la carte des chromosomes et le Genome Browser. Chaque élément ou groupe d'éléments présent dans la base de données possède une page descriptive.

Toutes les données présentes sur le site Génolevures sont librement disponibles et les instructions pour les citer y figurent. Le site web Génolevures est développé en utilisant une architecture "transfert d'état representationel" [6] (REST) et les URLs pour les ressources identifiées individuellement peuvent être construites systématiquement, par exemple les ressources concernant des éléments chromosomaux tel que les gènes (prefix/elt/Abbrev/Element_identifier) ou les familles de protéines (prefix/fam/Family_identifier).

Acknowledgements

Ce travail est co-financé par le CNRS (GDR 2354), l'ANR (ANR-05-BLAN-0331; GENARISE), la région Aquitaine ('Pôle de Recherche en Informatique') (2005-1306001AB, partiel) et par l'ACI IMPBIO (IMPB114, 'Génolevures En Ligne').

References

[1] Souciet JL, Génolevures Consortium. Special issue: Génolevures. FEBS Lett. 487:1–149, 2000.

[2] Sherman DJ, Durrens P, Beyne E, Nikolski M, Souciet JL. Genolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts. Nucleic Acids Res. 32:D315–D318F. , 2004 .

[3] Dujon B, Sherman DJ, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuvéglise C, Talla E, et al. Genome evolution in yeasts. Nature 430:35–44,2004.

[4] Sherman D, Durrens P, Iragne F, Beyne E, Nikolski M, Souciet JL. Genolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts. Nucleic Acids Res. 34:D432–D435, 2006

[5] Sherman DJ, Martin T, Nikolski M, Cayla C, Souciet JL, Durrens P. Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genome, Nucleic Acids Research, 37(Database issue):D550-D554, 2009

[6] Fielding R, Taylor RN. Principled design of the modern web architecture. ACM Trans. Internet Techn. 2:115–150, 2002.

HeliaGene : portail bioinformatique “tournesol”

Thibaut Hourlier, David Rengel, Nicolas Langlade, Patrick Vincourt,
Jérôme Gouzy, Sébastien Carrere

Laboratoire des Interactions Plantes Micro-organismes, UMR CNRS-INRA, F-31320 Castanet Tolosan

Thibaut.Hourlier@toulouse.inra.fr
Sebastien.Carrere@toulouse.inra.fr
Jerome.Gouzy@toulouse.inra.fr

Abstract: *A bioinformatics portal, called HeliaGene (http://www.heliagene.org) has been developed for in-depth analyses of Helianthus sp. EST data. This portal provides a variety of pre-computed analyses and tools for EST clusters and for exploring gene function and protein families in a user-friendly fashion. HeliaGene provides the biologists with an interactive access to the annotation of tens of thousands of clusters and their corresponding peptides as well as the bioinformaticians with a programmatic access to BioMoby web-services and BioMoby-based workflows.*

Keywords: Sunflower Genomics, Web-services, Workflows.

1 Introduction

Grâce d’une part à sa capacité d’adaptation aux environnements pauvres en eau des régions du sud de l’Europe et d’autre part à son potentiel de production de matériel pour les bio-carburants, l’espèce de tournesol *Helianthus annuus* est amenée à occuper une place de plus en plus importante parmi les plantes cultivées pour la production de biocarburants de première génération. Le projet de séquençage du génome commence à peine mais une quantité conséquente d’EST de sept espèces *Helianthus* est dès à présent disponible dans les banques publiques (284,251 au 18/01/2008).

Afin de permettre l’exploitation de ces premières ressources de séquences nous avons développé le portail HeliaGene dont le système de navigation permet (i) de rapidement visualiser les caractéristiques des clusters d’EST, (ii) d’explorer les fonctions des gènes, (iii) d’analyser les gènes et les familles de protéines, (iv) de rechercher des SNP potentiels à partir de polymorphisme intra et inter espèces.

Ainsi, la première analyse des données a été la prédiction de régions codantes à partir des clusters d'EST ; prédiction délicate du fait de l'hétérogénéité de ces derniers en terme de profondeur de couverture mais aussi par le polymorphisme induit par l’utilisation des séquences de sept variétés pour la génération des séquences consensus. Cet assemblage « multi-espèces » a été utilisé comme matrice de la puce affymetrix « tournesol » et sert donc de référence à la communauté. Pour prédire les peptides nous avons appliqué le programme FrameDP [1] qui est particulièrement adapté à la prédiction de CDS à partir de données bruitées. InterProScan a ensuite été utilisé pour analyser l'ensemble des peptides prédits afin d'en déterminer la composition en domaines et sites fonctionnels et pour proposer une classification basée sur la « Gene Ontology ».

2 Le portail http://www.HeliaGene.org

Le portail web http://www.heliagene.org/ est écrit en Perl-CGI. L'accès aux données se fait via des formulaires de recherche multicritères ou via des outils standard tels Blast ou PatScan. Une fiche synthétique reporte les caractéristiques de chaque entrée (clusters d'EST et peptides prédits) ainsi qu'un résumé des différentes analyses. Plusieurs niveaux d'authentification sont possibles, garantissant la confidentialité des données privées. Le portail propose aussi bien des interfaces pour les utilisateurs biologistes que des web-services BioMoby pour un accès programmatique aux données et aux programmes d'analyses.

2.1 Entrepôt de Données

Les données et les analyses sont structurées au format XML. Nous utilisons le moteur d'indexation CLucene (implémentation C++ de Lucene) afin d'indexer et stocker les fichiers XML. Pour ce faire, nous avons développé une suite de programmes Perl qui permettent à partir d'une description au format XPath d'indexer n'importe quel fichier XML, de requêter ces index et de générer un site web dynamique intégrant un formulaire de recherche multicritères. Nous avons packagé ces scripts sous le nom de EZLucene et étendu les fonctionnalités afin d'être capable d'analyser n'importe quel fichier texte à partir d'une description basée sur des expressions régulières Perl. Les résultats d'analyses bruts sont quant à eux compressés et stockés dans une BerkeleyDB afin de limiter le nombre et la taille des fichiers présents sur le système de fichiers tout en permettant un accès direct aux fichiers de données. Enfin, un système de moteur de template (AnotherTemplate.pm) permet un rendu HTML des fichiers XML.

2.2 Web Services et Workflows

L'ensemble des données et des programmes sont mis à disposition via des web-services ou des workflows BioMOBY[2], les web-services sont soumis à des procédures de test quotidiennes pour garantir leurs disponibilités. Ils sont de plus accessibles via le portail Mobyle[3] du LIPM ou à partir des gestionnaires de workflows tels Remora[4] et Taverna[5].

3 Disponibilité

HeliaGene est accessible à l'adresse http://www.heliagene.org. EZLucene, PlayMOBY et le serveur Mobyle du LIPM sont disponibles à partir de http://lipm-bioinfo.toulouse.inra.fr/.

References

[1] J. Gouzy, S. Carrere, T. Schiex. FrameDP: sensitive peptide detection on noisy matured sequences. *Bioinformatics*. 25:670-1, 2009.
[2] BioMoby Consortium. Interoperability with Moby 1.0--it's better than sharing your toothbrush! *Brief Bioinform.*, 3:220-31, 2008.
[3] B. Néron, H. Ménager, C. Maufrais, N. Joly, P. Tufféry, C. Letondal, Mobyle: a new full web bioinformatics framework., *Bio Open Source Conference* (BOSC), Toronto, 2008
[4] S. Carrere and J. Gouzy. REMORA: a pilot in the ocean of BioMoby web-services., *Bioinformatics*. 22:900-1, 2006.
[5] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M.R. Pocock, P. Li, T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.* W729-32, 2006.

Structural variants among transposable element families

Timothée Flutre [1], Hadi Quesneville [1]

[1] Unité de Recherche en Génomique-Info, UR 1164 INRA,
Route de Saint-Cyr, 78026 Versailles France
{Timothee.Flutre,Hadi.Quesneville}@versailles.inra.fr

Transposable elements (TEs) are repeated genomic sequences almost ubiquitous among prokaryote and eukaryote genomes and have a large impact on genome evolution. They are acknowledged as main agents involved in genome structure dynamics such as genome size variations and chromosomal rearrangements [1] but they can also be viewed as "controlling" elements [2] involved in epigenetics mechanisms [3] and the tinkering of gene regulatory networks [4].

As the number of sequencing projects is ever increasing, from model species to less studied ones, automatic *de novo* approaches are required to overcome the challenge of detecting nested, fragmented TEs in large, newly sequenced genomes. In this aim, we compared several programs and implemented a combined approach, the TEdenovo pipeline [5], now part of the REPET framework [6]. This comparative analysis has been performed on the *Drosophila melanogaster* release 4 genome as the Berkeley Drosophila Genome Project (BDGP) provides an exhaustive set of experimentally verified TEs present in this genome.

Our *de novo* approach, namely the TEdenovo pipeline, returns a set of classified, non-redundant consensus, each of them built from a multiple alignment of at least 3 genomic sequences. On the *Drosophila melanogaster* release 4 genome, we obtain 704 consensus, a number much larger than the 126 sequences present in the BDGP reference databank. We investigated this discrepancy and discovered that the classified, non-redundant consensus obtained with TEdenovo correspond to sub-families rather than families within the studied genome. We then reconstructed the families via an all-by-all comparison of the consensus followed by a clustering with a 50% coverage constraint. We then focused on several families by first looking at the multiple alignments of the consensus, and second at the multiple alignments of the genomic sequences used to make these consensus. While the former gave a hint, the latter showed that a TE family cannot be reduced to a single reference sequence because of the structural variants of the sub-families.

The fact that *de novo* approaches detect structural variants for each TE family should improve the subsequent annotation process [7]. Therefore, using several consensus for the same TE family, each of them corresponding to a specific structural variant, will not only benefit the annotation process but will also give insights into the dynamics of the genome subject to TE proliferation. Indeed, this lead us to analyse the patterns of indels present in the TE copies. Two statistical analyses are performed [8,9] in order to estimate the deletion rate. The first approach uses the positive, monotonic correlation between nucleotide substitutions and short deletions to estimate the deletion rate with a Poisson law. The other uses a pair-HMM to annotate the indels in a multiple alignment, and thus estimate the deletion rate. Therefore, these analyses allow us to quantify the rate at which the genome gets rid of TEs by means of small deletions.

Acknowledgements

This work is supported by the INRA.

References

[1] Eichler, E. E. & Sankoff, D. (2003), 'Structural dynamics of eukaryotic chromosome evolution.', Science 301(5634), 793—797.
[2] McClintock, B. (1956), 'Controlling elements and the gene.', Cold Spring Harbor symposia on quantitative biology 21, 197—216.
[3] Slotkin, R. K. & Martienssen, R. (2007), 'Transposable elements and the epigenetic regulation of the genome', Nature Reviews Genetics 8(4), 272—285.
[4] Feschotte, C. (2008), 'Transposable elements and the evolution of regulatory networks.', Nature Reviews Genetics 9, 397—405.
[5] Flutre, T., Duprat, E. & Quesneville, H., 'A comparative analysis of de novo detection methods of transposable elements in sequenced genome', in prep.
[6] the REPET framework, http://urgi.versailles.inra.fr/development/repet
[7] Buisine, N.; Quesneville, H. & Colot, V. (2008), 'Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets.', Genomics.
[8] Petrov, D. A.; Lozovskaya, E. R. & Hartl, D. L. (1996), 'High intrinsic rate of DNA loss in Drosophila.', Nature 384(6607), 346—349.
[9] Kim, J. & Sinha, S. (2007), 'Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment', Bioinformatics 23(3), 289—297.

Legoo: une plateforme bioinformatique pour la biologie intégrative des légumineuses

Marion Verdenaud[1]; Sébastien Carrere[1]; *Sebastien Letort*[1]; Emeline Deleury[2] , *Erika Sallet*[1]; Emmanuel Courcelle[1]; Olivier Stahl[3]; Thomas Faraut[3]; Vincent Savois[4]; Karine Gallardo[4]; Frédéric Debelle[1]; Pascal Gamas[1]; Jérôme Gouzy[1]

[1] Laboratoire des Interactions Plantes Micro-organismes (LIPM), UMR CNRS-, F-31320 Auzeville
[2] Interactions biotiques en santé végétale, INRA/CNRS/Univ. Nice, F-06900 Sophia Antipolis
[3] Laboratoire Génétique Cellulaire UMR444, INRA/ENVT, F-31320 Castanet Tolosan
[4] Unité de Rech. en Génétique et Ecophysiologie des Légumineuses à Graines, INRA, F-21110 Bretenières

Marion.Verdenaud@toulouse.inra.fr
Jerome.Gouzy@toulouse.inra.fr

Abstract: *Legoo is a bioinformatics portal (http://www.legoo.org) dedicated to legume studies. It targets knowledge integration and focuses on i) comparative analysis of genome sequences ii) integration of genome sequences with on one hand genetic maps of crops species and on the other hand transcriptomic data available on legume models iii) the structuring and representation of knowledge derived from the transcriptomic and proteomic approaches.*

Keywords: Genomics, comparative genomics, integration, knowledge base

1 Introduction

La biologie intégrative végétale peut être définie comme une intégration multidimentionnelle. Ainsi il convient, dans un premier temps, d'intégrer différents types de données « haut-débits ». Le second axe d'intégration doit permettre l'interprétation des données moléculaires à des niveaux d'organisations plus importants (cellule, tissu, etc.). La troisième dimension de l'intégration correspondant au transfert de connaissances entre les espèces modèles et d'intérêts agronomiques rendu possible par la comparaison de cartes génétiques et de génomes. « Legoo » fournit de nombreuses ressources le long des trois axes de l'intégration. Contrairement aux ressources existantes, Legoo se focalise sur la gestion et la représentation des connaissances à partir des réseaux de gènes et de comparaisons inter-espèces, proposant ainsi de multiple points de vues sur les données et les connaissances acquises chez les légumineuses.

2 Principales fonctionnalités du portail

2.1 Outils génomiques génériques.

Legoo fournit un accès aux données des légumineuses pour lesquelles des ressources génomiques ont été produites (*Medicago truncatula, Lotus japonicus,* soja). Ainsi une vingtaine de jeux de données peut être analysée à partir de serveurs « Blast » et « PatScan ». Les génomes annotés sont

quand à eux accessibles sous forme de fichiers plats et à partir des logiciels d'annotation `Artemis` [1] et `Apollo` [2] exécutés en mode web-start afin de simplifier leur utilisation.

2.2 Structuration et représentation des connaissances

L'objectif actuel est d'être capable de modéliser les processus biologiques pour tenter de prédire *in silico* l'évolution de ces processus sous différents contraintes. Afin d'entraîner ces modèles, il convient dans un premier temps de structurer et de collecter les connaissances. C'est pourquoi, nous avons implémenté une base de connaissance « minimale » qui va permettre de structurer sous forme de graphe les connaissances issues de la littérature ou d'une analyse experte. Une interface web permet de renseigner, d'interroger et de naviguer dans ce graphe de relation qui peut être également chargé automatiquement dans `cytoscape` [3].

2.3 Intégration des « omics » et transfert des connaissances entre espèces.

L'exploration du transcriptome de *Mt* a été initié bien avant l'achèvement de la séquence du génome, ce qui a conduit a utiliser dans la littérature de multiples nomenclatures liées aux différentes technologies et rend difficile la comparaison de listes de gènes. Nous avons donc développé l'outil « `nickname` » qui à partir de n'importe quel identifiant d'abord identifier le jeu de données auquel il appartient puis tous ses synonymes dans les jeux de données cibles.

Le mécanisme d'intégration des données proposé par `nickname` est complété par l'exploitation systématique des comparaisons de génomes via la plateforme `Narcisse` dédiée aux plantes [4]. Ainsi, `Narcisse` permet d'identifier et de représenter intéractivement les régions synténiques entre les différentes légumineuses. Il est ainsi aisé d'identifier les régions synténiques entre la plante modèle (Mt) et l'espèce d'intérêt agronomique, le soja (Gm), puis, pour une région synténique d'intérêt d'interroger dynamiquement la base de connaissance « *Medicago truncatula* » pour mettre graphiquement en vis-à-vis les données d'expression obtenues sur le modèle avec les gènes orthologues potentiels de l'espèce d'intérêt et ainsi faciliter le transfert de connaissances.

3 Disponibilité

`Legoo` est mis à disposition http://www.legoo.org aussi bien pour une utilisation par les utilisateurs finaux que sont les biologistes que pour un accès programmatique a partir des web-services BioMoby développés à partir du « framework » `PlayMOBY` (http://lipm-bioinfo.toulouse.inra.fr).

Remerciements

Le projet `Legoo` est en grande partie financé par l'ANR Genoplante 2006 GPLA06026G.

References

[1] T. Carver, M. Berriman, A. Tivey, C. Patel, U. Bohme, B.G. Barrell, J. Parkhill and M.A. Rajandream (2008) *Bioinformatics*, **24,** 2672-2676.
[2] S. Misra and N. Harris (2006) *Curr Protoc Bioinformatics*, **Chapter 9,** Unit 9 5.
[3] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker (2003) *Genome Res*, **13,** 2498-2504.
[4] E. Courcelle, Y. Beausse, S. Letort, O. Stahl, R. Fremez, C. Ngom-Bru, J. Gouzy and T. Faraut (2008) *Nucleic Acids Res*, **36,** D485-490.

ELIXIR : European Life Sciences Infrastructure For Biological Information

Antoine de Daruvar[1], Sandrine Palcy[1], Andrew Lyall[2], Janet Thornton[2]

[1] Centre de Bioinformatique de Bordeaux, Université de Bordeaux, Bordeaux, France
[2] European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, United Kingdom

Abstract: *ELIXIR is an EU Framework 7 Preparatory phase project for research infrastructures. Its goal is to construct and operate a sustainable infrastructure for biological information in Europe to support life science research and its translation to medicine and the environment, the bio-industries and society.*

Keywords: Infrastructure, databases, Biological Information.

1 Introduction

The objective of ELIXIR is to secure funding commitments from government agencies, research councils, funding bodies and scientific organisations within Europe, with the purpose of constructing a world-class and globally positioned European infrastructure for the management and integration of information in the life sciences.

2 Mission

To build a sustainable European infrastructure for biological information supporting life science research and its translation to medicine, the environment, the bioindustries, and society.

3 Benefits

ELIXIR will contribute to European science by:

- Optimising access to and exploitation of life-science data.
- Ensuring longevity of data and protecting investments already made in research.
- Increasing the competence and size of the user community by strengthening national efforts in training and outreach.
- Enhancing the global success and influence of Europe in life-science research and industry.

4 Rationale

To thrive, Europe needs:

- Coordinated data resources for the life sciences, with improved access and links with data in other related domains.
- A united European voice to influence global decisions and maintain open access.
- Adequate, sustainable funding for this distributed infrastructure

Acknowledgements

The European Life Science Infrastructure for Biological Information (ELIXIR) project is funded from the European Commission's Framework 7 Capacities Programme for Research Infrastructures under grant agreement no. 211601

Programmatic access to thousands of pre-computed transcriptional signatures using RTools4TB.

Fabrice Lopez[1,2], Aurélie Bergon[1,2], Julien Textoris[1,2,3], Jean Imbert[1,2], Samuel Granjeaud[1,2] and Denis Puthier[1,2,4]

[1] Inserm U928, TAGC, Parc Scientifique de Luminy case 928, 13288 Marseille Cedex 09, France
[2] Université de la Méditerranée, F-13007, Marseille, France
[3] Service d'Anesthésie et de Réanimation, hôpital Nord - Assistance Publique, Hôpitaux de Marseille, chemin des Bourelly, 13015 Marseille
[4] ESIL, Université de Provence et de la Méditerranée, 163 Avenue de Luminy, Case 925, 13288 Marseille Cedex 09, France

Abstract: *TranscriptomeBrowser is a software for integration and analysis of high-throughput genomic data. It is based on a database that stores a large collection of transcriptional signatures. This paper describe the development of new tools, a web service and the RTools4TB R package, that are intended to ease programmatic access to the database.*

Keywords: TranscriptomeBrowser, Transcriptome, Affymetrix, DBF-MCL, data mining.

1 Introduction

We recently developed a novel clustering algorithm, DBF-MCL ("Density Based Filtering and Markov Clustering"), that relies on nearest neighbor call analysis and on subsequent graph partitioning step using the Markov clustering [1][2]. One interesting aspect of DBF-MCL is that it was designed to handle noisy microarrays datasets as it can detect informative genes (those that fall into a cluster) prior to classification procedure. Taking advantage of the capabilities of DBF-MCL we searched clusters of co-regulated genes in a large panel of human, mouse and rat Affymetrix microarray datasets stored in the Gene Expression Omnibus database. All transcriptional signatures (TS) where stored in a relational database and a JAVA interface, TranscriptomeBrowser (TBrowser, http://tagc.univ-mrs.fr/tbrowser), was developed [1]. As reported earlier, TBrowser can be used to search though hundreds of experiments for the joint regulation of several genes or to find the biological contexts in which they are regulated. In order to ease programmatic access to the TranscriptomeBrowser (TBrowserDB) database we have (i) extended the capabilities of the search engine, (ii) developped a dedicated web service and (iii) build RTools4TB, an R package that implements the DBF-MCL algorithm and can perform calls to the web service.

2 TBrowser search engine improvement and web service development.

TranscriptomeBrowser comes with a sophisticated seach engine that allows complexe queries to be performed by the use of logical operators ("&", "|", "!"). As an example, user can search for TS containing the CD3E and CD4 markers but not the CD22 or CD14 marker :"CD3E & CD4 & !(CD19

|C14) ". We implemented a new search method that allows one to find TS containing at least a given proportion (*e.g.*, 80%) of a user-defined gene list. This novel type of query is particularly interesting when one wants to compare large gene lists (as those provided by all high throughput methods) to previously obtained microarray results. Moreover, still with the motivation of facilitating access to TBrowserDB, we developped a web service. Queries can be performed through the RTools4TB R package described below.

3 RTools4TB

RTools4TB (http://tagc.univ-mrs.fr/tbrowser/Rlib) is implemented in R programming language. For the representation of DBF-MCL results (DBFMCLresult class), we used the 'S4' system of formal classes and methods, that was popularized by the bioconductor project [3]. The core subroutine of DBF-MCL algorithm were written in C and are linked dynamically into R. Currently, the partitioning step is performed using a system call to the MCL application. This limits the use of RTools4TB to unix-like platforms. RTools4TB implements several popular normalization methods that can be applied to the dataset prior to classification (normal score transformation, quantile normalization, rank normalization). Furthermore, the DBF-MCL function can be used with various metrics for distance calculation (Euclidean distance, Pearson's correlation coefficient-based distance, Spearman's rank correlation-based distance). Finally, access to TBrowserDB can be done using various functions in order to retrieve a set of transcriptional signatures (based on gene composition, experiment, platform or annotation) or to retrieve informations about a microarray experiment or a platform.

4 Conclusions and prospects

We have developed several new tools to facilitate programmatic access to the TBrowserDB. In the future, we plan (i) to provide local database support and (ii) to integrate additional technologies such as ChIP-chip and ChIP-seq data.

Acknowledgements

The authors would like to thanks the staff from the TAGC laboratory for helpful discussions and gratefully acknowledge Francois-Xavier Theodule for technical assistance. This work was supported by the Institut National de la Santé et de la Recherche Médicale (Inserm), the Canceropole PACA and Marseille-Nice Genopole. Fabrice Lopez was supported by a fellowship from the EU STREP grant Diamonds and through funding from the IntegraTCell project (ANR, National Research Agency).

References

[1] Lopez F.,Textoris J., Bergon A., Didier G., Remy E., Granjeaud S., Imbert J. , Nguyen C. and Puthier D. TranscriptomeBrowser: a powerful and flexible toolbox to explore productively the transcriptional landscape of the Gene Expression Omnibus database. *PLoSONE*, 2008;3(12):e4001.

[2] Van Dongen S. (2000) A cluster algorithm for graphs. *National Research Institute for Mathematics and Computer Science* in the 1386-3681.

[3] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S *et al.*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80.

Narcisse, une représentation en miroir des synténies conservées

Sébastien Letort[1], Emmanuel Courcelle[1], Olivier Stahl[2], Jérôme Gouzy[1] et Thomas Faraut[2]

[1] Laboratoire des Interactions Plantes Micro-organismes, UMR 441-2594 (INRA-CNRS)
Jerome.Gouzy@toulouse.inra.fr
[2] Laboratoire de génétique cellulaire, UMR 444 (INRA-ENVT)
Thomas.Faraut@toulouse.inra.fr
INRA Toulouse, BP 52627, Chemin de Borde Rouge, Auzeville, 31326 Castanet Tolosan

Abstract: *New methods and tools are needed to exploit the unprecedented source of information made available by the completed and ongoing whole genome sequencing projects. The Narcisse database is dedicated to the study of genome conservation, from sequence similarities to conserved chromosomal segments or conserved syntenies, for a large number of animals, plants and bacterial completely sequenced genomes. The query interface, a comparative genome browser, enables to navigate between genome dotplots, comparative maps and sequence alignments.*

Keywords: Comparative genomics, completely sequenced genomes.

1 Introduction

L'un des défis majeurs de la bioinformatique de la période actuelle réside certainement dans sa capacité à exploiter pleinement l'information apportée par les génomes entièrement séquencés. Nous proposons d'accéder aux génomes séquencés et à leurs annotations à travers un navigateur de génomes comparés nommé Narcisse. Le nom de l'outil est inspiré par le principe adopté d'une représentation en miroir des segments chromosomiques conservés : les synténies conservées. L'originalité de l'outil réside, à nos yeux, dans sa capacité à caractériser différents niveaux de conservation, de l'alignement de séquences à la conservation de segments chromosomiques. Il est disponible à l'adresse suivante : http ://narcisse.toulouse.inra.fr [1].

2 Données et méthodes

Données Tous les génomes entièrement séquencés et publiquement disponibles sont intégrés à Narcisse. La grande majorité des génomes et leurs annotations sont téléchargés à partir de la division *genome* de genbank. Les génomes de certaines espèces de plantes sont téléchargés à partir de sites dédiés. Les données sont structurées en quatre "règnes" : bactéries, champignons, plantes et animaux. Seules les espèces d'un même règne sont comparées entre elles.

En plus du site web, nous proposons aux développeurs de workflows un accès aux données via des services web de type Biomoby [4].

Reconstruction des synténies conservées Au sein d'un même règne, toutes les comparaisons 2-à-2 de génomes sont réalisées aussi bien au niveau nucléique que protéique. Pour la comparaison nucléique, un programme nommé *glint*, développé par deux d'entre nous (Faraut et Courcelle, en préparation), permet une comparaison efficace grâce à un principe d'indexation de génomes. Le programme blast [2] est utilisé pour la comparaison protéique. Les alignements protéiques sont projetés sur le génome et combinés aux alignements nucléiques avant de procéder à l'identification de segments conservés. Les clusters d'alignements locaux qui présentent un certain degré de colinéarité entre les deux génomes sont considérés comme reflètant une conservation ancestrale de régions sous-jacentes. Afin de distinguer les réarrangements à petite échelle des réarrangements impliquant de larges régions chromosomiques, les conservations de synténie sont organisées sur plusieurs niveaux d'organisation hiérarchique.

Méthodes de visualisation Le navigateur de génomes comparés Narcisse permet d'explorer les niveaux de conservation à l'aide de différentes méthodes de représentation : dotplot, cartes comparées et représentations circulaires (Circos [3]). Des mesures quantitatives caractérisant les régions chromosomiques (gc%, densité en gènes, ...) complètent les annotations qualitatives habituelles (gènes, exon, rna, ...). Des fonctionalités avancées de zoom sont offertes, ainsi que la possibilité de choisir l'un ou l'autre des niveaux de conservation proposés, suivant le niveau de détail souhaité. Enfin, des liens hypertextes sont proposés vers les entrées correspondantes de Genbank.

Notre site permet également de comparer un génome de référence à plusieurs génomes cibles simultanément, aussi bien dans la représentation principale qu'avec la représentation circulaire Circos. Nous proposons également une comparaison d'un génome avec lui-même afin de localiser les duplications internes (en particulier pour l'instance dédiée aux plantes).

Acknowledgements

Nous souhaitons remercier l'Agence Nationale de la Recherche, GPLA06026G ANR Genoplante et ArcAnge ANR Genanimal, pour ses financemenets.

Références

[1] Courcelle E, Beausse Y, Letort S, Stahl O, Fremez R, Ngom-Bru C, Gouzy J, Faraut T (2008). Narcisse : a mirror view of conserved syntenies. *Nucleic Acids Res.*, 36 :D485-90.

[2] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1996). Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Res* 25 :3389-402.

[3] Krzywinski M. http ://mkweb.bcgsc.ca/circos/.

[4] Carrere S, Gouzy J. (2006) REMORA : a pilot in the ocean of BioMoby web-services. *Bioinformatics* 22 :900-1.

Hierarchical study of Guyton Circulatory Model

Rodrigo Assar Cuevas[1], Hayssam Soueidan[1], David J. Sherman[1]

INRIA Team MAGNOME and CNRS UMR 5800 LABRI, Université Bordeaux
351, cours de la Libération, F-33405 Talence France.
rodrigo.assar@labri.fr, hayssam.soueidan@labri.fr, david.sherman@inria.fr

Abstract: *This article presents an initial study of the Guyton Circulatory Model using BioRica. This model consists of* 18 *connected modules, each of which caracterise a separate physiological subsystem. We have focused the present analysis in the Renin-Angiotensin-Aldosterone System ($RAAS$). The use of BioRica allowed us to build an hierarchical model for this system by means of directly mapping modules to BioRica nodes. The results of each node were validated by comparison with published results.*

Keywords: Hierarchical Models, Blood Circulation, Renin-Angiotensin-Aldosterone System.

The Guyton model ([1]) is an extensive mathematical model of human circulatory physiology, that caracterises relations between conditions and physiological responses. Initially it defined relations between cardiac ouput and central venous and right atrial pressure, and was extended over time to include many physiological control processes. The Guyton Circulatory Regulation model consists of 18 connected modules, each of which characterises a physiological subsystem. Circulation Dynamics is the primary module. This model is naturally hierarchical, but historically has been defined using a flat collection of differential equations. We started analyzing the main points of the renal control of the blood pressure, in particular the Renin-angiotensin-aldosterone system ($RAAS$). The RAAS is crucial for the model ([2], [3]) and therapeutic manipulation of this pathway is very important in treating hypertension and heart failure. In general terms, a sustained fall in blood pressure causes the kidneys to secrete renin. This hormone stimulates the production of angiotensin in the circulation, whicht causes blood vessels to constrict, and stimulates the adrenal gland to produce aldosterone. This causes the tubules of the kidneys to retain sodium and water resulting in increased blood pressure. The mechanism of autonomic nervous controls of salt and water balance by the ADH (antidiuretic hormone) is also included. In this study we obtain, implement and test a true hierarchical model of the RAAS using BioRica. Our final goal is to build a more extensive model.

BioRica ([4]) is a high-level modeling framework developed by the MAGNOME team of INRIA that extends AltaRica ([5]) for biological applications, integrating discrete, continuous, stochastic, non deterministic and timed behaviors in a non-ambiguous way. The main advantage of using BioRica is that allows the caracterisation of hierarchical relations between nodes by means of dataflow links, leading to more expressive designs that can be more easily understood. Each module of the Guyton model that is associated to the RAAS was mapped into a BioRica node. We obtained the BioRica RAAS model, the hierarchical set of nodes and input-ouput relations between them. The nodes that were coded are: Angiotensin control, Aldosterone control, Antidiuretic hormone control (ADH mechanism), Electrolytes and cell water, Thirst and drinking and Kidney.

Each one of these nodes was either directly implemented in BioRica, or by encapsulating Matlab scripts that are used like SBML simulators. The implementations of each node were validated

by comparing its simulations results with results from the Physiome Project (see [6]), which were obtained using JSim (http://www.physiome.org/jsim/models/cellml/). Statistical analyses, Student's t-test (a parametric hypothesis test) and Kolmogorov-Smirnov for comparing two distributions (a nonparametric hypothesis test) were used to demonstrate that the values in simulation results are not significantly different. The parameters that are external to the $RAAS$ were fixed according to the default values of the Physiome Project.

The challenge is how to test the integration of the nodes into the hierarchical model. The first step was analyzing the physiological interpretation of the nodes of the BioRica $RAAS$ model. According to biological knowledge the system is activated when there is a loss of blood volume. Specialized cells (macula densa) of distal tubules sense the amount of sodium and chloride ions in the tubular fluid, and if it is low then renin is secreted, stimulating the production of Angiotensin. This process is represented in *Angiotensin node*. Angiotensin causes the secretion of Aldosterone, its production and functions is represented in *Aldosterone node*. Angiotensin also produces the blood vessels constrict, resulting in increased blood pressure, control that is represented by the relation ANM (multiplier effect of angiotensin), $MDFLW$ (rate of flow of fluid in the renal tubules at the macula densa) between *Angiotensin* and *Kidney node*. Aldosterone causes the tubules of the kidneys to retain sodium and water controlling the blood pressure. The *Kidney* inputs AMK (multiplier effect for control of potassium transport trough cell membranes) and $AMNA$ (multiplier effect for control of sodium) generated by *Aldosterone node* are used to compute the control of Na concentration, excretion of potassium and urine production by means of the outputs NOD (Na reabsorption), KOD (K secretion) and VUD (volume of urine). Another control process corresponds to the secretion of vasopressin, antidiuretic hormone (ADH) that promoves the reabsorption of fluid in the kidneys. The secretion of Antidiuretic hormone is represented by *Antidiuretic node*, linked with *Kidney node*. The production of vasopressin induces the reabsorption of water in the kidneys (*Thirst node*). In *Electrolytes node* is computed the concentration of K and Na by means of volume of fluid (TVD and VUD coefficients), rate of reabsorption of sodium NOD and rate of secretion of potassium KOD.

The second step was studying simulations. The direct effects of Angiotensin and the functions of Aldosterone and Antidiuretic hormone were reviewed by means of the inspection of the equations, relations and simulations. To test the results of the model we checked the control of the rate of flow of fluid in the renal tubules sensed by the macula densa, as one hoped it is taken to its normal level when initial level is low or high.

The doctoral thesis of Rodrigo Assar Cuevas is supported by INRIA. BioRica development is partially supported by YSBN EU FP6 LSHG 2005-18942.

References

[1] A.C. Guyton, T.G. Coleman and H.J. Granger, Circulation: Overall Regulation. *Annual Review of Physiology*, 34:13-44, 1972.

[2] A.C. Guyton, T.G. Coleman, A.W. Cowley, Jr., R.D. Mannings, Jr., R.A. Norman, Jr. and J.D. Ferguson, Brief Reviews: A Systems Analysis Approach to Understanding Long-Range Arterial Blood Pressure Control and Hypertension. *Circ. Res.*, 35:159-176, 1974.

[3] K. Sagawa, Critique of a Large-Scale Organ System Model: Guytonian Cardiovascular Model. *Annals of Biomedical Engineering*, 3:385-400, 1975.

[4] H. Soueidan, D.J. Sherman and M. Nikolski, BioRica: A multi model description and simulation system. *Foundations of Systems Biology and Engineering (F0SBE)*, 279-287, 2007.

[5] A. Arnold, G. Point, A. Griffault and A. Rauzy, The AltaRica Formalism for Describing Concurrent Systems. *Fundamenta Informaticae*, 34:109-124, 2000.

[6] J.B. Bassingthwaighte, Strategies for the Physiome Project. *Annals of Biomedical Engineering*, 28:1043-1058, 2005.

HasSium: a bioinformatic tool for fast reliable sorting and classification of very large samples of pyrosequenced amplicons

Aurélie Nicolas[1], Stéphane Avner[1], Alexis Dufresne[2], Stéphane Mahé[2], Philippe Vandenkoornhuyse[2], Frédérique Barloy-Hubler[1]

[1] Equipe B@SIC, UMR 6026 CNRS,
Université de Rennes 1, Campus Beaulieu 35042 Rennes France
aur.nicolas@aliceadsl.fr, savner@univ-rennes1.fr, fhubler@univ-rennes1.fr

[2] CNRS UMR 6553 ; ECOBIO
Université de Rennes 1, Campus Beaulieu 35042 Rennes France
Alexis.Dufresne@univ-rennes1.fr Stephane.Mahe@univ-rennes1.fr
Philippe.Vandenkoornhuyse@univ-rennes1.fr

Abstract: *HasSium is a free software developed to sort, count and classify a large set of data from pyrosequencing. HasSium help biologists studying molecular biodiversity by using multiple alignments and phylogenetic trees for sorting the set of sequences.*

Keywords: Sequence classification, Amplicon, Pyrosequencing.

1 Introduction

Since the development of pyrosequencing, biologists can approach new issues, like the biodiversity of ecosystems, the haplotype frequencies or allele quantification in a population and the mutation/SNP analysis. Biodiversity is often studied using amplicon analysis of one or several known genes in populations or samples. The pyrosequencing technology generates large amounts (many millions of sequences) of "noisy" short data due to imperfect reading accuracy and quality, and a lot of redundancy that needs to be counted. The assembling of numerous pyrosequencing reads into consensus sequences and the analyses of these sequences (clustering, classification, phylogeny, etc.) need the development of fast, dedicated and highly accurate bioinformatic tools. Indeed, already existing assembling and multiple sequence alignment algorithms are slow and/or innacurate. Specific tools were developed but suffer several limits [1, 2, 3]. In order to tackle these issues, a new application, HasSium, has been designed so as to be capable of addressing large amounts of sequences of the same gene, dealing with noisy data, sorting sequences into groups, assigning a kingdom, phylum or organism to every group, and finally counting and determining the diversity.

2 Method

2.1 HasSium Overall Pipeline

Since alignment algorithms have difficulties in dealing with millions of sequences, HasSium proposes to pre-filter the incoming data so that multi-sequence alignments will only be performed on datasets of reasonable size. These alignments will in turn be used to generate phylogenetic trees, which may help discriminate among sequences. Searching known anchors in sequences may then be employed to assign a kingdom, phylum or organism to every group of sequences. The project

finally involves counting the number of sequences and evaluating the diversity within every class. A diagram showing the different steps of the functional pipeline of HasSium is shown on Figure 1.

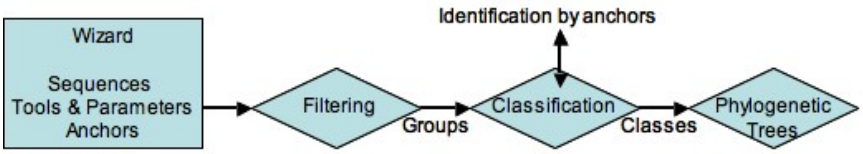


Figure 1. HasSium succesive steps.

2.2 Filtering Step

The first step of our algorithm consists in reducing the number of sequences that will be presented to the multiple sequence alignments algorithm. In order to sort the sequences using discrete criteria such as size, nucleotidic, di-nucleotidic and tri-nucleotidic content, boundaries will be homogenized using anchors specific of amplicon primers (when available) or amplified gene conserved sequences. Thus, at the end of the process, only sequences possessing the same size, nucleotidic, di-nucleotidic and tri-nucleotidic content will be collected in the same group. Finally, all sequences inside one group are compared to validate the homogeneity of the group. All sequences in one group are then considered to be similar.

2.3 Classification Step and Phylogenetic trees

At the end of the filtering step, one sequence per group is collected for classification. Anchors characterizing the kingdom, phylum and organism are implemented in HasSium for RNA 18S and RNA 16S. Biologists may also enter their own anchors and associated labels. The representative sequences collected by HasSium are classified using these anchors. After this step, multi-sequence alignments and phylogenetic trees are generated.

3 User Interface and Results

HasSium is a Java standalone application that makes use of a wizard to guide the biologist step by step, in order to produce a suitable set of parameters necessary to process the input data (the set of sequences). The parameters include the choice of primers, the identification anchors, the alignment algorithm, the phylogenetic tree algorithm. The results are presented in a graphical user interface that allows the biologist to show, parse and save images, sequences and tables for further analysis.

References

[1] F. Angly, B. Felts, M. Breitbart, P. Salamon, R. Edwards, et al. The marine viromes of four oceanic regions. PLoS Biol 4(11): e368, 2006.

[2] L. Krause, NN. Diaz, A. Goesmann, S. Kelley, TW. Nattkemper, F. Rohwer, RA. Edwards, J. Stoye. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res., 36(7):2230-2239, 2008.

[3] ML. Sogin, HG. Morrison, JA. Huber, D. Mark Welch, SM. Huse, PR. Neal, JM. Arrieta, GJ. Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. USA, 103(32):12115-12120, 2006

ProticWorkShop : un environnement bioinformatique pour la validation, l'analyse et l'intégration des données protéomiques*

Raphaël Flores[4], Laurent Gil[1], Daniel Jacob[1], Antoine de Daruvar[1], Delphine Vincent[2], Céline Lalanne[2], Christophe Plomion[2], Dominique Jeannin[3], Mireille Faurobert[3], Jean-Paul Bouchet[3], Benoît Valot[4], Michel Zivy[4], Olivier Langella[4] and Johann Joets[4]

[1] CBiB - Université Victor Segalen Bordeaux 2, 146, rue Léo Saignat, 33076 Bordeaux
[2] UMR BioGeCo, INRA, 69, route d'Arcachon, F-33612 Cestas Cedex
[3] UR GAFL, INRA, Domaine Saint-Maurice, BP 94, F-84143 Montfavet cedex
[4] UMR de Génétique Végétale, INRA, Univ Paris-Sud, CNRS, AgroParisTech, F-91190 Gif-sur-Yvette
joets@moulon.inra.fr

Abstract: *Due to a prominent automated production mode, proteomic data must be validated and/or curated before interpretation. This process, mostly manually performed, represents a bottleneck mainly resulting from a lack of suitable tools integrated into proteomic databases. Once validated, data often need to be subjected to various statistical analyses. Only few databases offer such functionality within their environment to date. In order to make these data accessible to as many scientists as possible, proteomic databases need to be included into international federative databases such as the "World-2DPAGE". To this end, we will develop PROTICws, a PROTICdb2-based bioinformatic environment dedicated to validation, analysis and sharing of proteomic data.*

Keywords: Database, comparative proteomics, mass spectrometry, two-dimensional polyacrylamide gel electrophoresis, Gene Ontology.

1 Introduction

L'approche la plus commune pour l'analyse du protéome est la combinaison de l'électrophorèse à deux dimensions (2-DE) pour la séparation des protéines et de la spectrométrie de masse (MS) pour leur identification. Elle permet de déchiffrer les processus biologiques dans le but de leur assigner une fonction [1]. De nombreuses données de protéomique ont été accumulées dans les laboratoires. Mais leur exploitation est souvent incomplète du fait du manque d'outils pour assister les protéomiciens dans les tâches de validation et d'analyse des données. De plus, le manque de standardisation pour la désignation des fonctions des séquences biologiques, la description des échantillons et des conditions expérimentales est un autre obstacle à l'échange et à l'analyse croisée des données. Pourtant des formats standards et des ontologies sont développés grâce à des initiatives internationales telles que celle du Proteomics Standard Initiative (PSI). Certaines bases de données telles que make2D-DB II [2],[3] et le « ***World-2DPAGE*** »ainsi que PROTICdb [4] sont compatibles avec ces standards, ce qui doit permettre de les fédérer facilement. L'objectif du projet PROTICworkshop est de répondre à ces problématiques en développant autour de PROTICdb un environnement logiciel pour valider, analyser et partager des données de protéomique.

* Le projet PROTICws est financé par l'ANR, programme Génomique.

2 Présentation de ProticWorkShop

2.1 La base de données PROTICdb 2

PROTICws est basé sur la base de données PROTICdb 2 qui est maintenant disponible en téléchargement pour utilisation en environnement de production. La version 2 de PROTICdb apporte de nouvelles fonctionnalités telles que la visualisation simplifiée de gels (Gel Browser), la présentation des spectres annotés et des tables d'ions en respectant le standard MIAPE ainsi que l'exportation des données dans les formats standards (mzXML 2.1, mzData v1.05, PRIDE 2.1).

2.2 Les modules de ProticWorkShop

PROTICannotate est un logiciel qui permettra d'annoter et de valider les données issues de PROTICdb 2. Il sera également possible de supprimer et d'éditer les données. Un moteur d'annotation automatique sera aussi implémenté, dans le but d'enrichir l'annotation des séquences biologiques à partir d'autres bases de données externes par recherche de similiarité et de domaines (GeneOntology, SwissProt, Interpro etc.). Les processus d'annotation pourront utiliser des données locales ou distantes. Enfin les données pourront être exportées dans des formats ad hoc pour des dépôts publics, locaux ou des outils tiers.

Le principal objectif de PROTICstat sera de faciliter l'accès aux outils statistiques pour les biologistes non spécialistes. L'idée est de mettre à disposition des chercheurs, via une interface web simple, un catalogue de chaînes de traitements statistiques pré-définies basées sur R. Ces chaînes seront décrites dans un manifeste au format XML de façon à pouvoir être créées par un bioinformaticien et implantées dans PROTICstat. Les utilisateurs pourront donc facilement adapter PROTICstat à leurs besoins. En plus des données présentes dans PROTICdb, ProticStat pourra accéder à d'autres bases de données telle que MERY-B (métabolomique) ou à des fichiers.

PROTICport comprendra une API pour dialoguer avec PROTICdb et une API pour créer des web-services sans avoir besoin d'une connaissance préalable de la structure de la base de données. En utilisant une unique fonction ou une combinaison de fonctions, un utilisateur avec des compétences minimales en informatique sera capable de créer des web-services spécifiques à ses propres besoins. PROTICport doit permettre également d'intégrer PROTICdb dans le réseau fédératif de bases de données de protéomique « World-2DPAGE »maintenu par le SIB.

Références

[1] S. Chen, A. C. Harmon (2006) *Proteomics* 6, 5504-16

[2] K. Mostaguir, C. Hoogland, P. A. Binz, R. D. Appel (2003) *Proteomics* 3, 1441-4

[3] C. Hoogland, K. Mostaguir, J. C. Sanchez, D. F. Hochstrasser, R. D. Appel (2004) *Proteomics* 4, 2352-6

[4] H. Ferry-Dumazet, G. Houel, P. Montalent, L. Moreau, O. Langella, L. Negroni, D. Vincent, C. Lalanne, A. de Daruvar, C. Plomion, M. Zivy, J. Joets (2005) *Proteomics* 5, 2069-81

Prioritization of scientific abstracts for biomedical research

Jean-Fred Fontaine [1], Adriano Barbosa-Silva [2], Miguel A. Andrade-Navarro [3]

[1] Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10 44035 D-13125 Berlin Germany
jean-fred.fontaine@mdc-berlin.de

[2] Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10 44035 D-13125 Berlin Germany
adriano.barbosa@mdc-berlin.de

[3] Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10 44035 D-13125 Berlin Germany
miguel.andrade@mdc-berlin.de

Abstract: *The Medline database contains millions of records which can be queried with the PubMed interface. Using this keyword-based Boolean search engine shows limitations for very specific topics, as it is difficult for a non-expert user to find all of the most relevant keywords. Additionally, when searching for more general topics, the same approach may return numerous unranked references. Text mining tools could help scientists focus on relevant abstracts. We have created the MedlineRanker webserver which allows a ranking of Medline for any topic of interest without expert knowledge. Given few abstracts related to a topic, the program finds automatically the most discriminative words in comparison to a random selection. Then, by using word relative frequencies, other abstracts can be ranked by relevance, including not annotated recent publications. We show that our tool can be highly accurate and that it is able to process millions of abstracts in a practical amount of time. The MedlineRanker webserver is available at http://cbdm.mdc-berlin.de/tools/medlineranker.*

Keywords: Data mining, Text retrieval, Biomedical literature, Text Mining.

1 Introduction

The PubMed query interface, which uses keywords to retrieve related biomedical documents, returns a list of abstracts, which are not sorted by relevance. If a search for a general topic is performed, hundreds or thousands of records may be returned; in this case, interesting abstracts may be hidden to the user because of their bad position in the list of results. Furthermore, for a very specific biological field, non-expert users would not be able to provide all the relevant keywords for the query. To improve text retrieval by scientists, text mining tools have been developed that offer alternative ways to query and select abstracts from the Medline database.

Some tools allow focusing on specific topics and filter out abstracts that are not relevant to the topic of interest (1-3). However, a proper set of keywords, which may not be obvious for a non-expert user, is still required to query the database. Making a query to Medline without using keywords or without knowing a specific vocabulary or query language is also possible using various text mining methods (4-5). Yet, these methods use one single abstract or text paragraph which may not be the best sample for a whole biomedical field and the resulting list is expected to contain irrelevant abstracts. Few methods have proposed automatic extraction of relevant information from a set of abstracts representing a topic of interest, and the use of this information to return a list of records ranked by relevance.

2 Results and Discussion

Based on an already published fast algorithm (6), we have implemented the MedlineRanker webserver, which allows a flexible ranking in Medline for a topic of interest without expert knowledge. The user defines their topic of interest using their own set of abstracts, which can be just a few examples, and can run the analysis with default parameters. If the input contains at least 100 closely related abstracts, the program returns relevant abstracts from the recent bibliography with high precision (up to 99%). This was illustrated with a benchmark ranking various topics, including complex topics defined by inter-related concepts, and comparison to similar ressources (7). Manual validation of 200 abstracts selected by MedlineRanker showed the relevance of our method which can lead to very high positive predictive values (up to 90 or 99%). It can process thousands of abstracts from the Medline database in a few seconds, or millions a in few minutes. It is not limited to specific topics and can be useful for all scientists interested in ranking or retrieving relevant abstracts from the Medline database, including specific subsets from particular databases.

References

[1] C. Perez-Iratxeta, P. Bork and M.A. Andrade, XplorMed: a tool for exploring MEDLINE abstracts. Trends Biochem Sci, 26, 573-575, 2001.
[2] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Rynbeek and P. Stoehr, Protein annotation by EBIMed. Nat Biotechnol, 24, 902-903, 2006.
[3] M.S. Siadaty, J. Shu and W.A. Knaus, Relemed: sentence-level search engine with relevance score for the MEDLINE database of biomedical articles. BMC Med Inform Decis Mak, 7, 1, 2007.
[4] J. Lin and W.J. Wilbur, PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics, 8, 423, 2007.
[5] J. Lewis, S. Ossowski, J. Hicks, M. Errami and H.R. Garner, Text similarity: an alternative way to search MEDLINE. Bioinformatics, 22, 2298-2304, 2006.
[6] B.P. Suomela and M.A. Andrade, Ranking the whole MEDLINE database according to a large training set using text indexing. BMC Bioinformatics, 6, 75, 2005.
[7] G.L. Poulter, D.L. Rubin, R.B. Altman and C. Seoighe, MScanner: a classifier for retrieving Medline citations. BMC Bioinformatics, 9, 108, 2008.

MeRy- B : management and analysis of plant metabolomics profiles obtained from NMR

Hélène Ferry-Dumazet[1], Laurent Gil[1], Antoine de Daruvar[1], Daniel Jacob[1]

[1] Centre de Bioinformatique de Bordeaux, Génomique Fonctionnelle Bordeaux, Université Bordeaux 2
146, rue Léo Saignat 33076 Bordeaux Cedex France
Daniel.jacob@u-bordeaux2.fr

Abstract: *Thanks to the improvement of metabolomics analytical techniques, more and more profiles are generated. The exploitation of these data needs platforms for profiles management and metabolites identification. Different databases exist for the management of plant metabolome profiles, such as the Golm Metabolome Database, which provides public access to Gas Chromatography – Mass Spectrometry (GC-MS) spectra. The Human Metabolome Database is another example of available knowlegebase for the human metabolome. However, currently, in the context of plant metabolomics, no platform exists to manage and analyse Nuclear Magnetic Resonance (NMR) metabolomics experiment. To this end, we will present MeRy-B, the first plant metabolomic platform allowing the storage and display of NMR plant metabolomics profiles.*

Keywords: Metabolomics, NMR, plant, databases, ontologies.

1 Introduction

Dans une démarche de génomique fonctionnelle, l'étude du métabolome est un complément indispensable de celle du transcriptome et du protéome. Parmi les techniques permettant l'acquisition de profils métaboliques, la spectrométrie de masse couplée à une séparation par chromatographie, et la spectroscopie par RMN du proton sont deux méthodes complémentaires. Ces techniques génèrent une grande quantité de données (metadata, raw data, peak list, analytes). Pour exploiter ces données, différentes bases de données proposent le stockage de profils métabolomiques : Golm Metabolome Database [1], Human Metabolome Database (HMDB) [2]… Cependant, aucune plateforme ne propose de stocker et analyser les études métabolomiques par RMN dans le domaine des plantes.
Un tel environnement de stockage et d'analyse des données, issues de RMN chez les plantes, est offert par l'outil MeRy-B.

2 Présentation de MeRy-B

2.1 Choix des outils et méthodes

Le schéma de la base de données est basé sur le modèle ArMet [3] et utilise le système de gestion de base de données PostgreSQL Les formats de description de données s'appuient sur MSI (OBO). Le format des spectres RMN supporté par MeRy-B est JcampDX. L'application est interfacée pour le Web.

2.2 Les fonctionnalités

Lors de la description de l'expérience, nous avons choisi d'utiliser les standards et formats recommandés par l'initiative internationale « Metabolomics Standards Initiative » impliquant les ontologies telles que le Plant Ontology Consortium (POC) pour le tissu et l'organe, la Taxonomy du NCBI pour les espèces et une partie de Environment Ontology (EO). A cela s'ajoute le format pdf, moins contraignant, permettant une description plus complète des protocoles.
L'outil MeRy-B offre quatre types d'interrogations : i) par projet , permettant une vue globale du plan expérimental sous forme de tableaux synthétiques, donnant accès aux protocoles et incluant une visualisation par analyses statistiques ; ii) par recherche de spectres selon des critères tels que l'espèce, le tissu, … ; iii) par recherche de métabolites selon son nom ou ses caractéristiques spectrales ; iv) par construction d'une question complexe dans le but d'exporter les données vers d'autres outils statistiques.

La base de connaissance MeRy-B a pour perspective de s'enrichir en accumulant des données provenant de divers laboratoires. L'outil MeRy-B est accessible à l'adresse suivante : http://www.cbib.u-bordeaux2.fr/MERYB/index.php.

Acknowledgements

This work was initially supported by Genoplante Consortium.

References

[1] Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmüller E, Dörmann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D. GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics. 2005 Apr 15;21(8):1635-8.

[2] Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, Macinnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ, Querengesser L. HMDB: the Human Metabolome Database. Nucleic Acids Res. 2007 Jan;35(Database issue):D521-6.

[3] Jenkins H, Hardy N, Beckmann M, Draper J, Smith AR, Taylor J, Fiehn O, Goodacre R, Bino RJ, Hall R, Kopka J, Lane GA, Lange BM, Liu JR, Mendes P, Nikolau BJ, Oliver SG, Paton NW, Rhee S, Roessner-Tunali U, Saito K, Smedsgaard J, Sumner LW, Wang T, Walsh S, Wurtele ES, Kell DB. A proposed framework for the description of plant metabolomics experiments and their results. Nat Biotechnol. 2004 Dec;22(12):1601-6

Effects of curine and guattegaumerine, two natural bisbenzylisoquinoline from Isolona hexaloba, on P-glycoprotein (MDR1) mediated efflux and in silico docking analysis.

Jacques-Aurélien Sergent[1], Hilarion Mathouet[2], Christian Hulen[1,3],

Akim Elomri[2], Nour-Eddine Lomri[1]

[1] Université de Cergy-Pontoise, GRP2H-INSERM UMRS 893, F-95000 Cergy-Pontoise
[2] Université de Rouen, CNRS UMR 6014, C.O.B.R.A. - I.R.C.O.F.
UFR Médecine-Pharmacie, 22 Boulevard Gambetta, 76183 Rouen Cedex 1, France
[3] University of Rouen, LMDF EA 4312, IUT, 55, Rue St Germain, 27000 Evreux, France.

Plants are an extraordinary reservoir of molecules that exhibit various biological activities on a large amount of different organisms. Here, we present the effects of two bisbenzylisoquinoline, curine and guattegaumerine isolated from roots of *Isolona hexaloba*, on the multidrug efflux pump mdr1b which is the homologue of human MDR1 pump. The predicted binding of these two molecules to human MDR1 protein was performed by using an in silico approach.

One the resistance mecanisms of cancer cells to anticancer agents, such as alkaloid drugs, is the extrusion of these molecules by efflux pumps like the membrane P-glycoprotein (P-gp) protein encoded by MDR1 gene. We, therefore, tested the effects of curine (H127) and guattegaumerine (H128) on the efflux of rhodamine 123, a substrate of P-gp, using a drug resistant cancer cell line (HTCR), and compared them to the effect of the verapamil, an inhibitor of P-gp.

Curine and guattegaumerine reduce the flow rate of rhodamine by 40 and 20% respectively, while the verapamil reduces this rate by almost 90%. The simultaneous addition of curine and guattegaumerine shows additive effects with a 68% decrease in the rate of Rhodamine efflux (Fig.1).

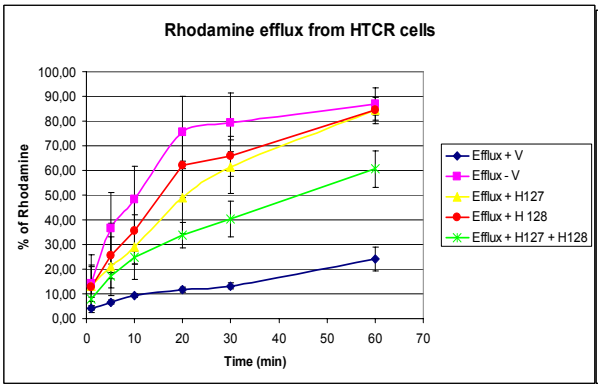


Figure 1. Rhodamine efflux from HTCR cells by P-gp

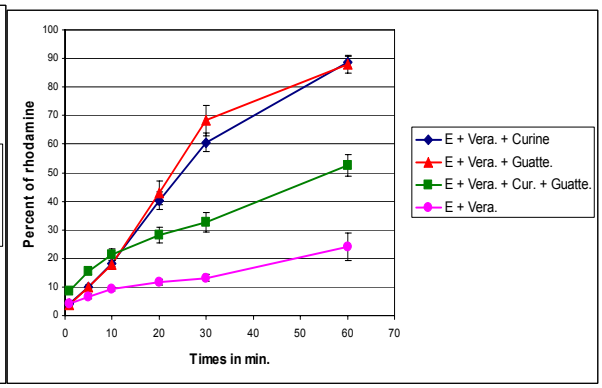


Figure 2. Rhodamine efflux from HTCR by P-gp in presence of verapamil and bibenzylisoquinolone

When curine or guattegaumerine is added to verapamil, inhibition of rhodamine efflux by verapamil disappears. Flow rates of rhodamine efflux are becoming identical to those observed with curine or guattegaumerine alone (Fig.2). The two bisbenzylisoquinolines seem to inhibit P-gp pump, and therefore, might compete with verapamil to alter R123 efflux.

To investigate any potential interaction between these two molecules and P-gp protein and to explain the inhibitory effect of these alkaloids, we performed docking analyses by using AutoDock 4.0.
To achieve this, we have used the available MDR1 PDB (1MV5 [3]) file and Openbabel to produce PDB files for all the chemicals used in theses studies (3D structure). Using AutoDock 4.0, we have identified interaction domains and binding sites for curine, guattegaumerine, verapamil and Rhodamine123 (fig 3) within MDR1 protein. The occupied positions by curine and guattegaumerine in MDR1 could explain the inhibitory effect observed on R123 efflux.
As an output of Autodock 4.0, Inhibitory Constants (Ki) were automatically estimated: 45nM for Curine, 137nM for Guattegaumerine and 1,28 µM for Verapamil. The graphical docking of curine or guattegaumerine interaction with P-gp protein is presented in figure 3. We can observe that curine competes directely with verapamil for the same binding whereas guattegaumerine interacts with P-gp on another site.

Figure 3. Docking of Curine (Volume) and Gauttegaumerine (Sticks) with p-GP
(arrows indicate verapamil binding site)

According to our findings, these two molecules might represent potential inhibitors of anti-cancer drugs efflux pumps.

References :

[1] Becker, J. P., G. Depret, F. Van Bambeke, P. M. Tulkens and M. Prevost (2009). "Molecular models of human P-glycoprotein in two different catalytic states." BMC Struct Biol 9(1): 3.
[2] Brown, R. S., Jr., N. Lomri, J. De Voss, C. M. Rahmaoui, M. H. Xie, T. Hua, S. D. Lidofsky and B. F. Scharschmidt (1995). "Enhanced secretion of glycocholic acid in a specially adapted cell line is associated with overexpression of apparently novel ATP-binding cassette proteins." Proc Natl Acad Sci U S A 92(12): 5421-5.
[3] Seigneuret, M. and A. Garnier-Suillerot (2003). "A structural model for the open conformation of the mdr1 P-glycoprotein based on the MsbA crystal structure." J Biol Chem 278(32): 30115-24.
[4] Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J. (1998). "Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function". J. Computational Chemistry, 19: 1639-1662.

High-throughput construction and optimization of 140 new genome-scale metabolic models

Christopher Henry[1], Matt DeJongh[2,3], Aaron Best[2], Paul Frybarger[2], Rick Stevens[1,4]

[1] Argonne National Laboratory, Argonne, IL 60439, USA
chrisshenry@gmail.com
[2] Hope College, Holland, MI 49423, USA
dejongh@hope.edu
best@hope.edu
paul.frybarger@hope.edu
[3] LaBRI, Université de Bordeaux 1, 33405 Talence cedex, France
[4] University of Chicago, Chicago, IL 60637, USA
stevens@anl.gov

Abstract: *Genome-scale metabolic models of organisms are useful for predicting genotypic and phenotypic characteristics, but the rate of model development has lagged far behind the rate of genome sequencing. We describe an automated process for genome-scale metabolic model generation that incorporates new developments in biochemical reaction network construction, and report the results of applying this process to generate models for 140 diverse prokaryotic genome sequences.*

Keywords: metabolic reaction networks, flux balance analysis, model optimization, prokaryotic genomes.

One of the valuable end products of the genome annotation process is the development of a genome-scale metabolic model of the organism being annotated [1]. These models provide a means of predicting genotypic and phenotypic characteristics of organisms, such as necessary growth conditions, gene essentiality, and response to genetic mutations, as well as potential for metabolic engineering [2]. However, the rate of genome-scale metabolic model development has lagged far behind the rate of genome sequencing for several reasons, including the presence of errors and gaps in the biochemical reaction databases used to build metabolic models, incomplete understanding of biochemical reaction directionalities, and errors and gaps in genome annotations [3].

Fortunately, solutions to these problems have emerged. Continuous refinement and expansion of biochemical databases such as the KEGG [4], and publication of numerous genome-scale metabolic models provide dependable resources for constructing reaction networks that represent much of the variety of central and intermediate metabolism. The recently developed Group Contribution method enables the prediction of reaction directionality based on thermodynamic feasibility [5]. New gap filling algorithms determine the minimal set of reactions that must be added to a metabolic network in order for every fundamental component of biomass to be synthesized from compounds available in the media [6]. The GrowMatch algorithm identifies how genome-scale metabolic

models must be modified in order to correct erroneous predictions regarding gene essentiality [7]. Finally, the SEED framework for comparative genome analysis enables the rapid, accurate annotation of newly sequenced prokaryotic genomes using subsystems technology [8].

We have extended the SEED framework to integrate these resources in the context of an automated process for construction of genome-scale metabolic models of prokaryotes [9]. We have applied this process to produce functioning models for a diverse set of 140 organisms across 14 bacterial divisions. On average, these models are comprised of 960 reactions associated with 624 genes covering 20% of the genome. Application of a gap filling algorithm results in the addition of an average of 83 reactions without gene associations. The GrowMatch algorithm is applied when gene essentiality and/or phenotyping data are available, resulting in some models that predict over 90% of the data. These extensions to the SEED framework are being incorporated into the RAST (Rapid Annotation based on Subsystems Technology) server [10], to enable the rapid construction and optimization of a genome-scale metabolic model for every sequenced prokaryotic genome.

Acknowledgements

Matt DeJongh, Aaron Best and Paul Frybarger are supported by the United States National Science Foundation MCB Award 0745100.

Matt DeJongh has received support from the Argonne National Laboratory Guest Faculty Program and an Aquitaine Regional Council – Fulbright Research Scholar Award

References

[1] A. Feist, M. Herrgard, I. Thiele, J. Reed, B. Palsson, Reconstruction of biochemical networks in organisms. Nat Rev Microbiol, 7(2):129-143, 2009.
[2] A. Feist, B. Palsson, The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat Biotechnol 26(6):659-667, 2008.
[3] C. Francke, R. Siezen, B. Teusink, Reconstructing the metabolic network of a bacterium from its genome. Trends Microbiol, 13(11):550-558, 2005.
[4] M. Kanehisa, S. Goto, S. Kawashima, A. Nakaya, The KEGG databases at GenomeNet. Nucleic Acids Res, 30(1):42-46, 2002.
[5] M. Jankowski, C. Henry, L. Broadbelt, V. Hatzimanikatis, Group contribution method for thermodynamic analysis of complex metabolic networks. Biophys J, 95(3):1487-1499, 2008.
[6] K. Satish, M. Dasika, C. Maranas, Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics, 8:212, 207.
[7] V. Kumar, C. Maranas, GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. PLoS Computational Biology, 5(3): e1000308, 2009.
[8] R. Overbeek, T. Begley, R. Butler, et al., The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res, 33(17):5691-5702, 2005.
[9] M. DeJongh, K. Formsma, P. Boillot, J. Gould, M. Rycenga, A. Best, Toward the automated generation of genome-scale metabolic models in the SEED. BMC Bioinformatics, 8:139, 2007.
[10] R. Aziz, D. Bartels, A. Best, M. DeJongh, et al., The RAST server: rapid annotations using subsystems technology. BMC Genomics, 9:75, 2008.

Protein Blocks: from simple structural approximation to multiple applications

Agnel Praveen Joseph[1], Aurélie Bornot[1], Bernard Offmann[2], Narayanaswamy Srinivasan[3], Manoj Tyagi[4], Hélène Valadié[5],Catherine Etchebest[1], Alexandre G. de Brevern[1]

[1] INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Paris Diderot - Paris 7, INTS, 6, rue Alexandre Cabanel 75739 Paris Cedex 15, France.
agnel.praveen@univ-paris-diderot.fr, aurelie.bornot@univ-paris-diderot.fr, Catherine.Etchebest@univ-paris-diderot.fr, alexandre.debrevern@univ-paris-diderot.fr

[2] INSERM UMR-S 665, DSIMB, Faculté des Sciences et Technologies, Université de La Réunion, 15 Avenue René Cassin, BP 7151, 97715 Saint Denis Messag Cedex 09,La Réunion, France.
bernard.offmann@univ-reunion.fr

[3] Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India.
ns@mbu.iisc.ernet.in

[4] Computational Biology Branch, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), 8600 Rockville Pike, Bethesda, MD 20894.
tyagim@ncbi.nlm.nih.gov

[5] UMR 5168 CNRS-CEA-INRA-Université Joseph Fourier, Institut de Recherches en Technologies et Sciences pour le Vivant, 17 avenue des Martyrs, 38054 Grenoble Cedex 9, France.
hvaladie@yahoo.fr

Abstract: *Protein structures are classically described in terms of secondary structures. Even if the regular secondary structures have relevant physical meaning, their recognition from atomic coordinates has some important limitations, and 50% of all residues are left undescribed, i.e. the coil. Thus different research teams have described local protein structures with the aim of approximating locally every part of the protein backbone. These libraries of local structures consist of sets of small prototypes named "structural alphabets". We have developed a structural alphabet, named Protein Blocks, not only to approximate the protein structure, but also to predict them from the sequence. Since its development, we and other teams have explored numerous new research fields using this structural alphabet. We review here some of the most interesting approaches.*

Keywords: Structural alphabet, protein structure, structure prediction, structural superimposition, mutation, binding site, Bayes theorem, Support Vector Machines.

The classical description of protein structures involves two regular states, the α-helices and the β-strands and one non-regular and variable state, the coil. Nonetheless, this simple definition of secondary structures masks numerous limitations. For instance, 3 states may over-simplify the description of protein structure; 50% of all residues, *i.e.*, the coil, are not described even if it encompasses repeating local protein structures. Description of local protein structures have hence focused on the elaboration of complete sets of small prototypes called "structural alphabets" (SAs),

that helps to analyze local protein structures and to approximate every part of the protein backbone [1]. The principle of a structural alphabet is simple. A set of average recurrent local protein structures is firstly designed. They approximate (efficiently) every part of known structures. As one residue is associated to one of these prototypes, we can translate the 3D information of the protein structures as a serie of prototypes (letters) in 1D, as the amino acid sequence. Our structural alphabet is composed of 16 structurally averaged protein fragments that are 5 residues length, called Protein Blocks (PBs, [2]). They have been used both to describe the 3D protein backbones and to perform a local structure prediction [3] (http://www.dsimb.inserm.fr/~debrevern/LOCPRED/). Our works on PBs have proven their efficiency in the description and the prediction of long fragments [4-6] and short loops [7], to define a reduced amino acid alphabet dedicated for mutation design [8], to analyze protein contacts [9] and in the building of a transmembrane protein. We have also used protein blocks to compare / superimpose protein structures (http://bioinformatics.univ-reunion.fr/PBE/). Developments made performed in other laboratories, using PBs, have focussed on the reconstruction of globular protein structures, the design of peptides, the definition of binding site signatures, novel prediction methodologies (http://www.fz-juelich.de/nic/cbb/service/service.php) and fragment-based local statistical potentials. The features of this alphabet have been compared by Karchin *et al.* with those of 8 other structural alphabets showing that our PB alphabet is highly informative, with the best predictive ability of those tested. It is nowadays the most widely used SA in the world. For a complete review of the different published SAs, please read ref [1].

Acknowledgements

These works were supported by grants from the Ministère de la Recherche, Université Paris Diderot – Paris 7, Université de Saint-Denis de la Réunion, the National Institute for Blood Transfusion (INTS) and the Institute for Health and Medical Care (INSERM). APJ has a grant from CEFIPRA number 3903-E, AB has a grant from the Ministère de la Recherche and HV has a post-doctoral fellowship from CEA. NS and AdB acknowledgement to CEFIPRA for collaborative grant.

References

[1] B. Offmann, M. Tyagi and A.G. de Brevern, Local Protein Structures, Current Bioinformatics, 3:165-202, 2007.
[2] A.G. de Brevern, C. Etchebest and S. Hazout, Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks, Proteins, 41:271-287, 2000.
[3] C. Etchebest, C. Benros, S. Hazout and A.G. de Brevern, A structural alphabet for local protein structures: Improved prediction methods, Proteins, 59:810-827, 2005.
[4] C. Benros, A.G. de Brevern, C. Etchebest and S. Hazout, Assessing a novel approach for predicting local 3D protein structures from sequence, Proteins, 62:865-880, 2006.
[5] C. Benros, A.G. de Brevern and S. Hazout, Analyzing the sequence-structure relationship of a library of local structural prototypes, J Theor Biol, 256:215-226, 2009.
[6] A. Bornot, C. Etchebest and A.G. de Brevern, A new prediction strategy for long local protein structures using an original description, Proteins, 2009, *in press*.
[7] L. Fourrier, C. Benros and A.G. de Brevern, Use of a structural alphabet for analysis of short loops connecting repetitive structures, BMC Bioinformatics, 5:58, 2004.
[8] C. Etchebest, C. Benros, A. Bornot, A.C. Camproux and A.G. de Brevern, A reduced amino acid alphabet for understanding and designing protein adaptation to mutation, Eur Biophys J, 36:1059-1069, 2007.
[9] G. Faure, A. Bornot and A.G. de Brevern, Analysis of protein contacts into Protein Units, 2009, *in press*.

Time specification in discrete models of biological systems

Nolwenn Le Meur[1,2], Michel Le Borgne[2], Jérémy Gruel[1,2], and Nathalie Théret[1]

[1] IRSET SeRAIC-INSERM, Université de Rennes I, rue du professeur Léon Bernard, Rennes Cedex, FRANCE
[2] IRISA Symbiose, Campus de Beaulieu, 35042 RENNES Cedex, FRANCE
nlemeur@irisa.fr

Abstract: *Modeling biological systems requires precise temporal concepts. Biological observations are often issued from discrete event measurements which make discrete modeling especially interesting. However time is absent of these models or defined a priori. In this paper, we propose a new formalism to specify time in discrete logical model and illustrate the power of our approach using a eukaryote cell cycle model.*

Keywords: System biology, discrete dynamical modeling.

1 Introduction

Among the formalisms available to model biological system, discrete modeling is especially appealing because the modeled events resemble to biological observations. Discrete model-ling is of particular interest as it benefits from a wealth of work done in circuit and program verification with well-established concepts and languages such as temporal logic, and efficient data structures and decision procedures. However, a major drawback of discrete model-ling, presently used in Systems biology, is the lack of specification on the order of the transitions. This is deferred to the interpretation of the model in a simulator. The common approaches are the synchronous and asynchronous interpretations. In synchronous mode, all the possible changes occur in one evolution step whereas in asynchronous mode, only one change is allowed at each step. Recently, authors [2] have introduced priorities on transitions to get finer description of time. Nevertheless, these assumptions are still not part of the model but directives to the simulator. With complex directives on the sequencing of transitions, analysis and verification methods rapidly become impossible to implement. Here, we propose a new formalism to include time inside the model and use a unique interpretation for all models. This formalism is inspired by the formal models underlying real time programming languages like Signal [1]. Time is the logical time used in computer science: it does not correspond to the duration of events but to their relative sequencing. This allows the description of several biological signals with different clocks, *i.e.*, multiclock systems.

2 Time in discrete model : A language for specifying multiclock system

To experiment with the multiclock concept applied to biological models, we are developing *Biosignal*, a language for specifying time in discrete logical models using. A multiclock dynamical system is specified by *states*, *signals* and *clock constraints*. For instance for the variable X:

```
state(X, r_X, false); clock(h_X)
```

```
r_X := ((not Y) and (not Z)) when h_X
synchro(h_X, h_Y)
```

`X` is the state variable, `r_X` is the refreshing signal that modify `X`, and `false` is the initial value. `clock(h_X)` is the free clock of `X`, *ie.*, a time parameter of the system. Clock constraint between the signals `h_X` and `h_Y` is synchronous but could be exclusive or ordered by priorities.

3 Results

Using the *Biosignal* language, we illustrate the use of clocks in the implementation of time specifications with the cell cycle model proposed by [2]. The simulation of our model with the synchronous and asynchronous specifications and identical initial state as in [2], give isomorphic transition graphs. These simulation results shows that our approach gives the same behaviors as an approach where timing considerations are in the interpretation of the model and not in the model. Using the same model and formalism, we can also apply formal verification. For instance, we studied the behavior of our model without any clock constraints. To this aims, we introduced Boolean constraints corresponding to the 4 successive phases of the cell cycle (i.e., $G1,\ S,\ G2,\ M$) and based on the activity/inactivity of the CDK-cyclin complexes as summarized in literature. We were then able to verify using model checking techniques that our model reaches S (whether or not starting in $G1$). We also showed that with specific clock constraints, the model can reach $G1$ then S without going through $G2 \cup M$.

4 Conclusions

We introduced a new formalism inspired by computer science formal models, to build models of biological systems based on discrete event systems. This formalism is implemented as a specification language: *Biosignal* . We showed that our formalism is well suited for building models with complex clock constraints. Compare to simulation approaches, we do not separate clock constraints from logical conditions. The first advantage of having both in the same formal model is to provide a clear semantic of clock constraints. The second advantage is to make clock constraints amenable for proofs. In particular, this allows the use of model checking, whatever the complexity of the clock constraints. Finally, it gives means to compute on models. This has an important potential for model fitting and experiment planning. Presently, *Biosignal* is a primitive language. We hope that practical use of *Biosignal* will foster specification patterns which could be turned into elaborate language constructions. Property specification for model checking will be also implemented.

Acknowledgements

This work is supported by the Institut National de la Santé et de la Recherche Médicale, the Ligue Contre le Cancer, and the Region Bretagne (PRIR 3193).

References

[1] A. Benveniste, P. Bournai, T. Gautier, M. Le Borgne, P. Le Guernic, and H. Marchand. The Signal declarative synchronous language: controller synthesis & systems/architecture design. *40th IEEE Conference on Decision and Control*, 2001.

[2] A. Faure, A. Naldi, C. Chaouiya and D. Thieffry, Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 14:e124-131, 2006.

EcoPrimer : a new program to infer barcode primers from full genome sequence analysis

Tiayyba Riaz[1], François Pompanon[1], Pierre Taberlet[1], Eric Coissac[1,2]

[1] Laboratoire d'Ecologie Alpine (LECA)
CNRS UMR 5553 2233, Université Joseph Fourrier, BP 53, 38041 Grenoble Cedex 9, France
tiayyba.Riaz@bvra.e.ujf-grenoble.fr
francois.pompanon@ujf-grenoble.fr
pierre.taberlet@ujf-grenoble.fr
[2] INRIA Rhône-Alpes – Projet Hélix
ZIRST-655 Avenue de l'Europe 38334 Montbonnot Cedex
eric.coissac@inrialpes.fr

Abstract: *Species identification by DNA barcoding has gained sufficient coverage today. In this paper we present an informatics program, EcoPrimer, to infer a barcode usable region and barcode primers allowing its PCR amplification from a set of sequences. To develop this program, we propose two indices to measure the quality of barcode region called specificity and coverage. The actual program works around finding the conserved regions from the input set of sequences to design PCR primers and to measure the quality of the region between the two conserved regions to use it as barcode.*

Keywords: DNA Barcoding, Primers, Specificity, Coverage.

Introduction DNA barcoding is a tool for characterizing the species origin using a short sequence from a standard position in the genome [**?**]. From a practical point of view, a barcode locus should be flanked by two conserved regions to design PCR "universal" primers. Several manually discovered barcode loci are routinely used today, but no objective function has been described to measure their quality in terms of universality (barcode coverage, B_c) or in terms of taxonomical discrimination capacity (barcode specificity, B_s). In this paper, we propose a more formal approach to qualify a barcode region (B_c , B_s functions) and a new way for identifying barcode loci without *a priori* on the candidate sequences. Our barcode inferring method is the combination of an algorithm to detect conserved regions from a set of full genome sequences to design PCR primers and objective functions to determine the quality of a barcode locus.

Materials and Methods 4000 sequences of fully sequenced mitochondrial genomes were extracted from genbank and filtered according to their taxa resulting in 837 unique taxon sequences which were used as training set. The *EcoPrimer* software is written in C language.

Results The measurement of the quality of barcode region is based on two factors; the *barcode specificity* (B_s) and *barcode coverage* (B_c).

Measure of the specificity of a barcode region The specificity of a barcode region measures the ability of this region to discriminate between two taxa. It is the ratio of well identified taxa to

the total number of taxa. The number of well identified taxa are calculated by putting in a relation a region and a set of taxon, via the individuals of the taxon and the barcodes that they own for this region. We say that a taxon t is well identified by a barcode region r if the barcode regions possessed by the individuals of the taxon are not shared by the individuals of any other taxon.

Measure of the coverage of a primer pair The coverage index is the ratio of amplified taxon to the total number of taxa in the input data set. This second measure aims to estimate the quality of the primers in terms of their capacity to amplify a broad range of species

Barcode Region Inference from whole genomes: ecoPrimer The software *ecoPrimer* is developed using a heuristic approach in which it first finds the strict primer which are present in at-least a certain percentage of sequences specified by quorum value. Then it locate approximate copies the strictly identified primers with the Baeza-Yates/Manber algorithm [**?**] (Agrep algorithm). Finally primers are paired by taking into consideration the distance between two primers on a sequence and quality is measured using the above mentioned indices.

ecoPrimer can also find the primers specific to a subclade of the input sequences by giving its taxid. The sequences which correspond to this taxid are called example sequences and the others are called counter example sequences. The software finds primers which are present in more than Q_e example sequences and less than Q_c counter example sequences. It is some times important to find the primers which are spcific to a particular subgroup of sequences. This is particularly important when you work on ancient DNA to reduce noise induce by contamination with modern DNA.

Some newly identified regions on birds In collaboration with the Museum d'Histoire Naturelle de Grenoble, we want to determine the origin of the feathers of an Egyptian mummy's pillow. The main problem is to analyse an ancient vertebrate DNA of bird, contaminated by human DNA another vertebrate. To achieve this, our program was run on the 837 vertebrates data set considering the 71 birds sequences as example sequences. The size of the barcode DNA was limited to 150 bp to take into account the degraded status of our DNA sample. The results shown in table 1 correspond to three primer pairs which potentially amplify a small region of the 16S RNA gene of 98% of these 71 species.

Primer Name	Sequences		Count	Family		Genus		Species		Fragment size (bp)		
	Direct	Reverse		B_s	B_c	B_s	B_c	B_s	B_c	min	max	average
P01	aaaaacatagccttcagc	gccattcatacaagtctc	68	94.87	97.5	89.29	98.25	84.62	98.48	98	108	101.3
P02	aaaaacatagccttcagc	agccattcatacaagtct	68	94.87	97.50	89.29	98.25	84.62	98.48	99	109	102.3
P03	aaaaacatagccttcagc	tagccattcatacaagtc	68	94.87	97.50	89.29	98.25	84.62	98.48	100	110	103.3

Table 1. A selection of three primer pairs proposed by our software for birds. The program was run with the following parameters : $primer_{size} = 18bp$, $quorum = 70\%$, $min_{length} = 20bp$, $max_{length} = 1000bp$.

References

[1] P.D.N. Hebert, E.H. Penton, J.M. Burns, D.H. Janzen and W. Hallwachs, Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proc Natl Acad Sci*, 101(41):14812-14817, 2004.

[2] S. Wu, U. Manber. Agrep – a fast approximate pattern-matching tool. *Proceedings USENIX Winter 1992 Technical Conference*, strony 153–162, San Francisco, CA, 1992.

Long range expression effects of copy number variation: insights from Smith-Magenis and Potocki-Lupski syndrome mouse models

Guénola Ricard[1], Jacqueline Chrast[1], Jessica Molina[2,3], Nele Gheldof[1], Sylvain Pradervand[1], Frédéric Schütz[1,4], James R. Lupski[5,6,7], Katherina Walz[2] and Alexandre Reymond[1]

[1]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland
[2]Centro de Estudios Científicos (CECS), Valdivia, Chile
[3]Universidad Austral de Chile, Valdivia, Chile,
[4]Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland
[5]Molecular & Human Genetics, Baylor College of Medicine, Houston, Texas, USA
[6]Pediatrics, Baylor College of Medicine, Houston, Texas, USA
[7]Texas Children's Hospital, Houston, Texas, USA.
guenola.ricard@unil.ch

Abstract: *To study the effect of structural changes on expression we assessed gene expression in genomic disorder mouse models. Both a microdeletion and its reciprocal microduplication mapping to mouse chromosome 11 (MMU11), which model the rearrangements present in Smith-Magenis (SMS) and Potocki-Lupski (PTLS) syndromes patients, respectively, have been engineered. We profiled the transcriptome of five different tissues affected in human patients in mice with 1n (Deletion/+), 2n (+/+), 3n (Duplication/+) and uniallelic 2n (Deletion/Duplication) copies of the same region in an otherwise identical genetic background. The most differentially expressed transcripts between the four studied genotypes were ranked. A highly significant propensity, are mapping to the engineered SMS/PTLS interval in the different tissues. A statistically significant overrepresentation of the genes mapping to the flanks of the engineered interval was also found in the top-ranked differentially expressed genes. A phenomenon efficient across multiple cell lineages and that extends along the entire length of the chromosome, tens of megabases from the breakpoints. These long-range effects are unidirectional and uncoupled from the number of copies of the CNV genes. Thus our results suggest that the assortment of genes mapping to a chromosome is not random. They also indicate that a structural change at a given position of the human genome may cause the same perturbation in particular pathways regardless of gene dosage. An issue that should be considered in appreciating the contribution of this class of variation to phenotypic features. We will also discuss the molecular networks that are altered in the different models. This network analysis enables the identification of metabolic pathways that potentially play a function in the SMS/PTLS phenotypes and allows a better comprehension of the roles of the different genes of the interval.*

Keywords: Copy Number Variation, Epigenetics, Transcriptomics, Diseases, Position effect, chromosome rearrangement.

Uncovering overlapping clusters in biological networks

Pierre Latouche, Etienne Birmelé, Christophe Ambroise

Laboratoire Statistique et Génome, UMR CNRS 8071-INRA 1152-UEVE
La Genopole, Tour Evry 2, 523 place des Terrasses 91000 Evry France
pierre.latouche@genopole.cnrs.fr

Abstract: *In the last few years, there has been a growing interest in studying biological networks. Many deterministic and probabilistic clustering methods have been developed. They aim at learning information from the presence or absence of links between pairs of vertices (genes or proteins). Given a network, almost all these techniques partition the vertices into disjoint clusters, according to their connection profile. However, recent studies have shown that these methods were too restrictive and that most of the existing biological networks contained overlapping clusters. To tackle this issue, we present in this paper a latent logistic model, that allows each vertex to belong to multiple clusters, as well as an efficient approximate inference procedure based on global and local variational techniques. We show the results that we obtained on a transcriptional regulatory network of yeast.*

Keywords: Biological networks, clustering methods, overlapping clusters, global and local variational approaches.

1 Model and Notations

We consider a directed binary random graph $\mathcal{G}$, where V denotes a set of N fixed vertices and $\mathbf{X} = \{X_{ij}, (i,j) \in \mathrm{V}^2\}$ is the set of all the random edges. We assume that $\mathcal{G}$ does not have any self loop, and therefore, the variables X_{ii} will not be taken into account.

For each vertex $i \in \mathrm{V}$, we introduce a latent vector $\mathbf{Z}_i$, of Q independent Boolean variables $Z_{iq} \in \{0,1\}$, drawn from Bernoulli distributions:

$$\mathbf{Z}_i \sim \prod_{q=1}^{Q} \mathcal{B}(Z_{iq};\, \alpha_q) = \prod_{q=1}^{Q} \alpha_q^{Z_{iq}} (1-\alpha_q)^{1-Z_{iq}}, \tag{1}$$

and we denote $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_Q\}$ the vector of class probabilities. Note that in the case of a usual mixture model, $\mathbf{Z}_i$ would be generated according to a multinational distribution with parameters $(1, \boldsymbol{\alpha})$. Therefore, the vector $\mathbf{Z}_i$ would see all its components set to zero except one such that $Z_{iq} = 1$ if vertex i belongs to class q. The model would then verify $\sum_{q=1}^{Q} Z_{iq} = \sum_{q=1}^{Q} \alpha_q = 1$, $\forall i$. In this paper, we relax these constraints using the product of Bernoulli distributions (1), allowing each vertex to belong to multiple classes. We point out that $\mathbf{Z}_i$ can also have all its components set to zero.

Given two latent vectors $\mathbf{Z}_i$ and $\mathbf{Z}_j$, we assume that the edge X_{ij} is drawn from a Bernoulli distribution:

$$X_{ij} | \, \mathbf{Z}_i, \mathbf{Z}_j \sim \mathcal{B}\big(X_{ij};\, \mathrm{g}(a_{\mathbf{Z}_i,\mathbf{Z}_j})\big) = e^{X_{ij} a_{\mathbf{Z}_i,\mathbf{Z}_j}} \mathrm{g}(-a_{\mathbf{Z}_i,\mathbf{Z}_j}),$$

where

$$a_{\mathbf{Z}_i,\mathbf{Z}_j} = \mathbf{Z}_i^\intercal \, \mathbf{W} \, \mathbf{Z}_j + \mathbf{Z}_i^\intercal \, \mathbf{U} + \mathbf{V}^\intercal \, \mathbf{Z}_j + W^*, \tag{2}$$

and $\mathrm{g}(x) = (1 + e^{-x})^{-1}$ is the logistic sigmoid function. $\mathbf{W}$ is a $Q \times Q$ matrix whereas $\mathbf{U}$ and $\mathbf{V}$ are Q-dimensional vectors. The first term in (2) describes the interactions between the vertices i and j. If i belongs only to class q and j only to class l, then only one interaction term remains ($\mathbf{Z}_i^\intercal \, \mathbf{W} \, \mathbf{Z}_j = W_{ql}$). However, the interactions can become much more complex if one or both of these two vertices belong to multiple classes. Note that the second term in (2) does not depend on $\mathbf{Z}_j$. It models the overall capacity of vertex i to connect to other vertices. By symmetry, the third term represents the global tendency of vertex j to receive and edge. Finally, we use W^* as a bias, to model sparsity.

2 Experiments

We consider the yeast transcriptional regulatory network described in Milo et al. (2002) and we focus on a subset of 192 vertices connected by 303 edges. Nodes of the network correspond to operons, and two operons are linked if one operon encodes a transcriptional factor that directly regulates the other operon. Such networks are known to be relatively sparse which makes them hard to analyze. In this Section, we aim at clustering the vertices according to their connection profile. Using $Q = 6$ clusters, we apply our algorithm and we obtain the results in Table 1.

cluster	size	operons
1	2	STE12 TEC1
2	33	YBR070C MID2 YEL033W SRD1 TSL1 RTS2 PRM5 YNL051W PST1 YJL142C **SSA4** YGR149W SPO12 YNL159C SFP1 YHR156C YPS1 YPL114W HTB2 MPT5 SRL1 DHH1 **TKL2** PGU1 YHL021C RTA1 WSC2 GAT4 YJL017W TOS11 YLR414C BNI5 YDL222C
3	2	MSN4 MSN2
4	32	CPH1 **TKL2** HSP12 SPS100 MDJ1 GRX1 SSA3 ALD2 GDH3 GRE3 HOR2 ALD3 SOD2 ARA1 HSP42 YNL077W **HSP78** GLK1 DOG2 HXK1 RAS2 **CTT1** HSP26 TPS1 TTR1 HSP104 GLO1 **SSA4** PNC1 MTC2 YGR086C **PGM2**
5	2	YAP1 SKN7
6	19	YMR318C **CTT1** TSA1 CYS3 ZWF1 HSP82 TRX2 GRE2 SOD1 AHP1 YNL134C **HSP78** CCP1 TAL1 DAK1 YDR453C TRR1 LYS20 **PGM2**

Table 1. Classfication of the operons into $Q = 6$ clusters. Operons in bold belong to multiple clusters.

First, we notice that the clusters 1, 3, and 5 contain only two operons each. These operons correspond to hubs which regulate respectively the nodes of clusters 2, 4, and 6. More precisely, the nodes of cluster 2 are regulated by STE12 and TEC1 which are both involved in the response to glucose limitation, nitrogen limitation and abundant fermentable carbon source. Similarly, MSN4 and MSN2 regulate the nodes of cluster 4 in response to different stress such as freezing, hydrostatic pressure, and heat acclimation. Finally, the nodes of cluster 6 are regulated by YAP1 and SKN7 in the presence of oxygen stimulus. In the case of sparse networks, one of the clusters often contains most of the vertices having weak connection profiles, and is therefore not meaningful. Conversely, with our approach, the vectors $\mathbf{Z}_i$ can have all their components set to zero, corresponding to vertices that do not belong to any cluster. Thus, we obtained 85 unclassified vertices. Our algorithm was able to uncover two overlapping clusters (operons in bold in Table. 1). Thus, SSA4 and TKL2 belong to cluster 2 and 4. Indeed, they are co-regulated by (STE12, TEC1) and (MSN4 and MSN2). Moreover, HSP78, CTT1, and PGM2 belong to cluster 4 and 6 since they are co-regulated by (MSN4, MSN2) and (YAP1, SKN7).

Une statistique de sphéricité pour l'adéquation d'un graphe à des données transcriptomiques

Vincent Guillemot[1,2], Arthur Tenenhaus[1] et Vincent Frouin[2]

[1] Département Signaux et Systèmes Électroniques, Supélec, 3, rue Joliot Curie, F-91190 Gif-sur-Yvette,
[2] Laboratoire d'Exploration Fonctionnelle des Génomes, CEA, 2, rue Gaston Crémieux, F-91000 Evry
vincent.guillemot@cea.fr

Abstract: *When analyzing genomic data, the researcher often encounters the situation where different genetic regulation graphs can be determined from the same dataset. One graph is often compared to another with the help of databases gathering already known regulations: the more known interactions an inferred graph contains, the better it is. We propose a different approach, adapted from the theory of sphericity tests, to determine whether a graph fits the dataset.*

Keywords: Genomics, graph, hypothesis test, sphericity test.

La détermination de graphes d'interactions génétiques est un problème récurrent en bioinformatique. Ces graphes peuvent être construits soit à partir de bases de données faisant l'inventaire de toutes les régulations observées dans des expériences de biologie, soit par des méthodes d'inférence comme la méthode *Graphical Lasso* (*glasso*) [5]. Ce genre de méthodes fait l'hypothèse qu'un profil d'expression est une réalisation d'une variable aléatoire $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. En se plaçant dans le cadre des modèles graphiques gaussiens [2], la matrice de précision $\mathbf{\Sigma}^{-1}$ de la variable X permet de déterminer un graphe de régulations génétiques.

Comment vérifier qu'un graphe est en adéquation avec un jeu de données transcriptomiques particulier ? Parmi plusieurs graphes, quel est celui qui est le plus en adéquation avec les données dont on dispose ? Les réponses que l'on apporte le plus souvent sont basées sur la comparaison des graphes obtenus avec un ou plusieurs graphes provenant de bases de données. La littérature propose des alternatives sous la forme de tests « locaux » d'adéquation d'un graphe à un jeu de données [3,2]. Nous proposons une approche globale pour déterminer si un graphe $\mathcal{G}$ donné correspond de façon significative à la matrice de précision empirique de données transcriptomiques. Nous utilisons pour cela une statistique inspirée de tests de sphéricité. Dans la suite, les données seront supposées gaussiennes multivariées.

Soit une variable aléatoire $\mathcal{N}(\mu, \mathbf{\Sigma})$. On considère un échantillon de cette variable, un individu étant noté $\mathbf{x}_\alpha, \alpha = 1, ..., (n+1)$. On pose de plus $\mathbf{S} = \frac{1}{n}\sum_{\alpha=1}^{n+1}(\mathbf{x}_\alpha - \overline{\mathbf{x}})(\mathbf{x}_\alpha - \overline{\mathbf{x}})'$.

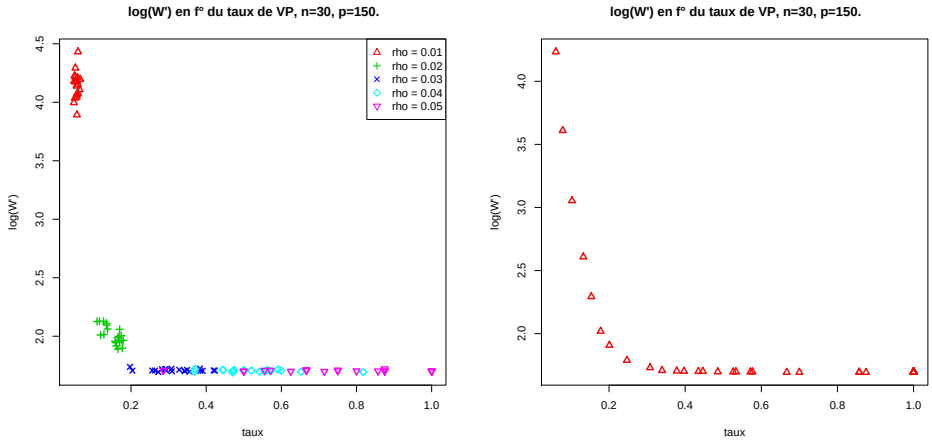
La statistique que nous proposons est notée W' :

$$W' = \frac{1}{p}\operatorname{tr}([\mathbf{S}\mathbf{\Sigma}_0^{-1} - I_p]^2) - \frac{p}{n}\left(\frac{1}{p}\operatorname{tr}(\mathbf{S}\mathbf{\Sigma}_0^{-1})\right)^2 + \frac{p}{n}.$$

W' est inspirée de la statistique W présentée dans [4], elle est généralisée au cas où $\mathbf{\Sigma}_0^{-1}$ n'est pas forcément égale à I_p grâce à une transformation proposée par [1]. Elle permet de comparer la matrice de covariance $\mathbf{\Sigma}_0$ déduite du graphe inféré $\mathcal{G}$ à $\mathbf{S}$. $\mathbf{\Sigma}_0^{-1}$ est déduite de $\mathcal{G}$ de la façon suivante : si $\mathbf{A}$ est la matrice d'adjacence de $\mathcal{G}$, $\mathbf{D}$ la matrice (diagonale) des degrés de $\mathcal{G}$ et I_p la matrice identité, alors $\mathbf{\Sigma}_0^{-1} = \mathbf{A} + \mathbf{D} + I_p$.

Les résultats que nous avons obtenus sont uniquement basés sur des données simulées, $n+1 = 30$ réalisations *i.i.d* d'une variable $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)$ à $p = 150$ composantes selon l'algorithme suivant : générer un graphe aléatoire $\mathcal{G}_{ref}$ de référence, générer un jeu de données $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{ref})$ tel que $\mathbf{\Sigma}_{ref} = (\mathbf{A}_{ref} + \mathbf{D}_{ref} + I_p)^{-1}$ et enfin obtenir des graphes avec la méthode *glasso* en variant le paramètre ρ de la méthode [5].

La figure 1(a) présente une comparaison entre la statistique W' et τ, le taux d'arêtes communes entre $\mathcal{G}_{ref}$ et $\mathcal{G}$. Chaque point correspond à la simulation de profils d'expression de taille 30×150, ρ variant entre $0,01$ et $0,05$. La figure 1(b) présente également W' en fonction de τ, mais pour un seul jeu de données correspondant à un seul graphe $\mathcal{G}_{ref}$. On applique ensuite la méthode *glasso* pour inférer des graphes avec différentes valeurs de ρ (prises entre $0,01$ et $0,2$).



(a) Pour des jeux de données indépendants. (b) Pour un même jeu de données.

Fig. 1. W' en fonction du taux d'arêtes correctement inférées par *glasso*. n et p sont fixés.

Les deux figures 1(a) et 1(b) permettent de constater que l'évolution de W' est décroissante en fonction de τ. De plus, lorsque deux graphes sont proposés pour un seul jeu de données, les statistiques obtenues sont tout à fait comparables et permettent de déterminer quel graphe est le plus en adéquation avec les données. La statistique W' présentée permet donc de comparer plusieurs graphes inférés sur un jeu de données transcriptomiques, ce classement ayant été validé sur des données simulées gaussiennes multivariées. Il faut maintenant travailler sur les résultats asymptotiques présentés par Ledoit et Wolf pour développer un test d'hypothèses.

Références

[1] T. W. Anderson. *An introduction to multivariate statistical analysis*, 3rd edition, Wiley, 2003.

[2] M. I. Jordan. *Learning in Graphical Models*, The MIT Press, 1998.

[3] N. Verzelen and Fanny Villers. Tests for gaussian graphical models. *CSDA, (to appear)*, 2008.

[4] O. Ledoit and M. Wolf. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Stat.*, 30 :1081-1102, 2002.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432-441, 2008.

Bi-dimensionnal Gaussian mixture for IP/IP ChIP-chip data analysis

Caroline Bérard[1], Marie-Laure Martin-Magniette[1,2], François Roudier[3], Vincent Colot[3] & Stéphane Robin[1]

[1] UMR AgroParisTech/INRA MIA 518
16 rue Claude Bernard, 75231 PARIS Cedex 05, France
caroline.berard@agroparistech.fr
marie_laure.martin@agroparistech.fr
stephane.robin@agroparistech.fr
[2] UMR INRA 1165 - CNRS 8114 - UEVE URGV
2 rue Gaston Crémieux, 91057 EVRY, France
[3] UMR CNRS 8186, Ecole Normale Supérieure, Département de Biologie
46 rue d'Ulm, 75230 PARIS Cedex 05, France
roudier@biologie.ens.fr
colot@biologie.ens.fr

Abstract: *Chromatin immunoprecipitation on chip (ChIP-chip) is a well-established procedure to investigate proteins associated with DNA. ChIP-chip enables to study differences between two immunoprecipitated DNA samples. From a biological point of view, we expect to distinguish four different groups: a group of non-immunoprecipited DNA, a group of immunoprecipited DNA in both samples, and then two groups in which DNA is immunoprecipited differently. We propose to model these data with a mixture of two-dimensional Gaussians with four components. Biological knowledges are included as constraints on the variance matrices. The parameters are estimated by the EM algorithm. This method is applied to NimbleGen data in order to study the histone methylation difference between the wild ecotype of model plant Arabidopsis thaliana and a mutant.*

Keywords: Gaussian mixture, Model selection, ChIP-chip

L'immunoprécipitation de la chromatine sur puce (ChIP-chip) est une technique utilisée pour étudier les interactions entre protéines et ADN. Habituellement dans une expérience de ChIP-chip, les deux échantillons co-hybridés sont les fragments d'ADN liés à la protéine d'intérêt (IP) et l'ADN génomique total (INPUT), mais le ChIP-chip permet également d'étudier directement la différence entre deux échantillons d'ADN immunoprécipité (issus d'un sauvage et d'un mutant par exemple), sans hybrider sur la puce l'ADN génomique total (INPUT). On s'attend alors à distinguer quatre groupes différents : un groupe d'ADN non-immunoprécipité, un groupe d'ADN immunoprécipité identiquement dans les deux échantillons, et puis deux groupes dans lesquels l'ADN est immunoprécipité différemment. L'objectif de notre travail est de proposer une modélisation conjointe des signaux IPs obtenus par un modèle de mélange de gaussiennes bi-dimensionnelles à quatre composantes, où les connaissances biologiques sont prises en compte sous forme de contraintes sur les paramètres du modèle.

Soit $X_i = (X_{1i}, X_{2i})$ le signal log-IP de chaque échantillon pour la sonde *i*, la densité du couple s'écrit :

$$f(X_i) = \sum_{k=1}^{4} \pi_k \phi(X_i|\mu_k, \Sigma_k),$$

où π_k est la proportion de la *k*ème composante du mélange ($0 < \pi_k < 1$, $\forall\ k = 1, ..., 4$ et $\sum_{k=1}^{4} \pi_k = 1$) et $\phi(.|\mu_k, \Sigma_k)$ est la densité d'une distribution gaussienne bidimensionnelle de paramètres (μ_k, Σ_k), où μ_k est la moyenne et Σ_k est la matrice de variance-covariance. Nous reprenons une paramétrisation proposée par [1] qui considère la décomposition spectrale des matrices de variance des classes. Cette paramétrisation permet de proposer de nombreux modèles de classification [2].
Afin d'intégrer la connaissance biologique, des contraintes sont ajoutées au modèle : le groupe d'ADN non-immunoprécipité et le groupe normal ont la même matrice d'orientation car ils sont tous les deux orientés selon la première bissectrice. D'autre part, on suppose que le bruit est égal dans chaque groupe, ce qui correspond à fixer la deuxième valeur propre de Σ_k. En effet, la première valeur propre est associée au premier axe de l'ellipse (grand axe) et la deuxième est associée au petit axe de l'ellipse. Nous obtenons ainsi :

$$\begin{cases} \Sigma_k = D_k \Lambda_k D_k', \text{ pour } k = 1, .., 4 \\ D_1 = D_2 = D \\ \Lambda_k = \begin{pmatrix} \lambda_{1k} & 0 \\ 0 & \lambda_2 \end{pmatrix}, \text{ pour } k = 1, .., 4, \text{ avec } \lambda_{1k} > \lambda_2 \end{cases}$$

où Λ_k représente le volume et la forme, D_k représente l'orientation. D_k est la matrice des vecteurs propres de Σ_k et Λ_k est une matrice diagonale avec les valeurs propres de Σ_k sur la diagonale dans l'ordre décroissant.
Les paramètres $(\pi_1, ..., \pi_3, \mu_1, ..., \mu_4, \Sigma_1, ..., \Sigma_4)$ sont estimés à l'aide d'un algorithme EM et les sondes sont classées dans l'un des 4 groupes selon la règle du Maximum A Posteriori.

Nous illustrons cette méthode avec des données issues de la technologie NimbleGen. Les deux échantillons co-hybridés sur la puce concernent la méthylation d'une histone de la plante modèle *Arabidopsis thaliana* pour un sauvage et un mutant.

Références

[1] J.D. Banfield and A.E. Raftery, Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49, 803-821, 1993.

[2] G. Celeux and G. Govaert, Gaussian Parsimonious Clustering Models. *Pattern Recognition* 28, 781-793, 1995.

OxyGene&Co : Combining OxyGene and CoBalt to improved the functional annotation of oxidative stress sub-systems

David Thybert[1]*, David Goudenège[1]*, Stéphane Avner[1], Céline Miganeh-Lucchetti[1] , Frédérique Barloy-Hubler[1]
* equal contribution to the work
[1] Equipe B@SIC, CNRS UMR 6026, Université de Rennes 1,
Campus de Beaulieu, Av. du Général leclerc, 35042 Rennes, France
{david. thybert, david. goudenege, stephane.avner, celine.lucchetti, fhubler}@univ-rennes1.fr

Abstract: *In order to improve fucntionnal annotation of prokaryotic genomes, we combine OxyGene and CoBalt in a new tool, OxyGene&Co. This tool add a subcellular localization layer to OxyGene oxidative stress annotation. OxyGene&Co demonstrate that different location can be predicted for proteins of the same detoxification subfamily.*

Keywords: Genomes annotation, oxydative stress, subcellular localization.

1 Introduction

Since the massive delivery of genome sequences, improving the annotation accuracy and homogeneity have become crucial since frequent imprecise, ambiguous or erroneous annotations [1] limit efficient comparative genomics and can cause interpretation mistakes regarding functional potentialities carried by genomes.

2 OxyGene and CoBalt

In order to deal with these problems, our group developed an innovative annotation strategy based on a new ontology organized in subsystems, *i.e* a set of gene classes and subclasses that share functions involved in the same biological process, and an anchor-based *ab-initio* genome annotation. This approach was implemented in a platform called OxyGene [3], used to annotate and analyze oxidative stress response subsystems in prokaryotic genomes. It improved the functional detection of ROS/RNS subsystem potentiality in prokaryotic genomes by simplifying and homogenizing the description of functions, by correcting errors and detecting forgotten loci [2].

In parallel, we also developed CoBalt (Consensus Based Localization Tool) [4] a database of the sub-cellular localization prediction the proteins. A very large number of specialized (signal peptide, alpha helices …) and global tools (localization prediction), using different methods (HMM, NN, SVM, patterns...), have been tested and 41 were selected. All the NCBI prokaryotic complete genomes were pre-computed and all protein results were stored in a specific interfaced database ,organized in prediction boxes, (lipoprotein, signal peptide, helical transmembrane...).

3 OxyGene&Co

As protein function is closely related to its activity location, OxyGene annotations have been associated with CoBalt predictions into a functional platform called OxyGene&Co. By combining OxyGene and Cobalt, we demonstrate that different subcellular localization can be predicted for proteins of the same ROS/RNS detoxification subfamily and that these location variations point out functional biodiversities, important for "genotype to phenotype" predictions.

Acknowledgements

This work is supported by Le Conseil Régional de Bretagne.

References

[1] JL. Bidartondo, Preserving accuracy in GenBank. Science 2008, 319(5870):1616.
[2] M. Diehn, G. Sherlock, G. Binkley, H. Jin, JC. Matese, T. Hernandez-Boussard, CA. Rees, JM. Cherry, D. Botstein, PO : Brown and others, SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data, Nucleic Acids Res 2003, 31(1):219-23
[3] D. Thybert, S. Avner, C. Lucchetti-Miganeh, A. Cheron, F. Barloy-Hubler : OxyGene: an innovative platform for investigating oxidative-response genes in whole prokaryotic genomes. BMC Genomics 2008, 9:637.
[4] D. Goudenège, S. Avner, F. Barloy-Hubler : COBALT – COmparaison of Bacterial Proteins Localisation Tools. JOBIM 2008.

The PSI Semantic Validators
How Compliant is Your Proteomics Data ?

Samuel Kerrien[1], Luisa Montecchi-Palazzi[1], Florian Reisinger[1],Bruno Aranda[1], Andrew R Jones[2], Matthias Oesterheld[3], Lennart Martens[1], and Henning Hermjakob[1].

[1] European Molecular Biology Laboratory (EMBL) – European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK
[2] Liverpool University Liverpool L69 3BX United Kingdom
[3] Institute for Bioinformatics, GSF National Research Center for Environment and Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany

skerrien@ebi.ac.uk

Abstract: *The advance of proteomics technologies has exposed the community to a deluge of data. The Proteomics Standard Initiative (PSI) has defined standards and controlled vocabularies (CVs) for the representation of experimental data. These data formats are commonly used and ensuring compliance to guidelines has become increasingly important.The PSI Semantic Validator addresses this issue by offering a framework for data validation, it allows the verification of correct CVs usage and complex integrity checking of the data. The framework is free, open-source and can be adapted to any data format.*

Keywords: Proteomics, Standardisation, Data validation.

1 Methods

The framework is written in Java and can be based on any other data model. Given an instance of a data type and a set of rules, the validator will process the data and return a set of messages, each of which reports on inconsistencies found. The error level of each message denotes its severity. Different types of validation can be performed such as syntax of the data (if provided in a textual form such as XML), correct usage of CVs (including consistent use of CV hierarchy) and advanced data check based on rules created by a validator's developer.

It is common for a mass spectrometry experiment to report very large amount of data (i.e. XML of a few gigabytes). The Validator was developed with this constraint in mind so that it can process large volume of data with minimal memory footprint.

2 Results

The Proteomics Standards for Molecular Interactions [1] (PSI-MI) and Mass Spectrometry (PSI-MS) have been widely adopted by the community and a large volume of experimental data is now available for scientists too (e.g. IntAct [2], PRIDE [3]). These data formats allow the effortless

aggregation of datasets, but falls short when it comes to enforcing complex constraints that ensure the data is biologically correct. The PSI Semantic Validator allows scientists to validate their proteomics data files as existing implementations of the validator have already been made available on the Internet for PSI-MI and PSI-MS (http://psidev.info/validator). Furthermore, the framework has been written such that it can be adapted to any data type, provided it is based on a well defined structure.

Another use of the validator is to check on user data submission to public repositories. Indeed, multiple rule sets can be developed for the same data format, thus defining different levels of stringency. For instance, data submission to the IntAct database could be checked against both the MIMIx guidelines [4], the IMEx curation manual [5] as well as the more stringent IntAct manual curation standards [6].

3 Innovative Aspects

- Generic framework adaptable to any data format.
- Seamless integration of CVs available in Open Biomedical Ontologies (OBO) format via the Ontology Lookup Service.
- Optimized handling of XML data so that it can process large volumes with ease.

References

[1] S. Kerrien, S. Orchard, L. Montecchi-Palazzi, B. Aranda, A. F. Quinn, N. Vinod, G. D. Bader, I. Xenarios, J. Wojcik, D. Sherman, M. Tyers, J. J. Salama, S. Moore, A. Ceol, A. Chatr-aryamontri, M. Oesterheld, V. Stümpflen, L. Salwinski, J. Nerothin, E. Cerami, M. E. Cusick, M. Vidal, M. Gilson, J. Armstrong, P. Woollard, C. Hogue, D. Eisenberg, G. Cesareni, R. Apweiler, H. Hermjakob. Broadening the Horizon – Level 2.5 of the HUPO-PSI Format for Molecular Interactions. BioMed Central. 2007.

[2] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, H. Hermjakob. IntAct – open source resource for molecular interaction data. Nucleic Acids Res. 2006 Dec 1; 17145710.

[3] L. Martens, H. Hermjakob, P. Jones, C. Taylor, K. Gevaert, J. Vandekerckhove, R. Apweiler. (2005) PRIDE: The PRoteomics IDEntifications database Proteomics Vol 5 Issue 13 Pages 3537-3545.

[4] S. Orchard, L. Salwinski, S. Kerrien, L. Montecchi-Palazzi, M. Oesterheld, V. Stümpflen, A. Ceol, A. Chatr-aryamontri, J. Armstrong, P. Woollard, J. J Salama, S. Moore , J. Wojcik, G. D Bader, M. Vidal, M. E Cusick, M. Gerstein, AC. Gavin, G. Superti-Furga, J. Greenblatt, J. Bader, P. Uetz, M. Tyers, P. Legrain, S. Fields, N. Mulder, M. Gilson, M. Niepmann, L. Burgoon, J. De Las Rivas, C. Prieto, V. M Perreau, C. Hogue, H. Werner Mewes, R. Apweiler, I. Xenarios, D. Eisenberg, G. Cesareni & H. Hermjakob. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nature Biotechnology 2007.

[5] S. Orchard, S. Kerrien, P. Jones, A. Ceol, A. Chatr-aryamontri, L. Salwinski, J. Nerothin, H. Hermjakob. Submit Your Data the IMEx way: a step by step Guide to Trouble-fee Deposition. Proteomics 2007.

[6] http://www.ebi.ac.uk/~intact/site/doc/IntActAnnotationRules.pdf

Comparison of Spectra in Unsequenced Species

Freddy Cliquet[1,2], Guillaume Fertin[1], Irena Rusu[1], Dominique Tessier[2]

[1] LINA, UMR CNRS 6241, 2 rue de la Houssinière, 44322, Nantes, Cedex 03, France
{freddy.cliquet,guillaume.fertin,irena.rusu}@univ-nantes.fr
[2] UR1268 BIA, INRA, Rue de la Géraudière, BP 71627, 44316 Nantes, France
dominique.tessier@nantes.inra.fr

Keywords: Proteomics, MS/MS, Spectra Alignment, Unsequenced Species.

1 Introduction

We introduce a new algorithm for the mass spectrometric identification of proteins. Experimental spectra obtained by tandem MS/MS are directly compared to theoretical spectra generated from proteins of evolutionarily closely related organisms. This work is motivated by the need of a method that allows the identification of proteins of unsequenced species against a database containing proteins of related organisms. The idea is that matching spectra of unknown peptides to very similar MS/MS spectra generated from this database of annotated proteins can lead to annotate unknown proteins. This process is similar to ortholog annotation in protein sequence databases. The difficulty with such an approach is that two similar peptides, even with just one modification (i.e. insertion, deletion or substitution of one or several amino acid(s)) between them, usually generate very dissimilar spectra. In this poster, we present a new dynamic programming based algorithm: PacketSpectralAlignment (PSA). Our algorithm is tolerant to modifications and fully exploits two important properties that are usually not considered: the notion of **inner symmetry**, a relation linking pairs of spectrum peaks, and the notion of **packet** inside each spectrum to keep related peaks together.

2 Results

We compare our algorithm PSA to SpectraAlignment [1,2] on a set of simulated data. We generate a dataset of 1000 random peptides of random size in $[10, 25]$ in order to constitute a database that will be used to create the theoretical spectra. Each peptide in the database is then modified by applying 0 to 4 random substitutions of amino acids to create simulated experimental spectra. Then we compare each simulated experimental spectrum with each theoretical spectrum.
Our tests show that the two algorithms have a comparable behaviour for 0 to 2 shifts, with a slight advantage for our algorithm. However, for more than two shifts, SpectralAlignment presents a fast deterioration of its results, while PacketSpectralAlignment still gives good results (see Figure 1).

We have also evaluated the benefits supplied by the packets (a packet is a cluster of several peaks). Thanks to packets, we do not test all masses in an experimental spectrum, but only those masses m inducing an alignment of at least T peaks when a packet from the theoretical spectrum is positionned at mass m. To evaluate this, we have computed the number of masses to treat for different values of T on four different datasets (one of simulated spectra and 3 of experimental maize spectra). Table 1 shows the evolution of the number of possible masses in function of the threshold T for each set of spectra. We can notice that the number of possible masses decreases considerably when T is increased. We also note, on the first tests (Fig. 1), with a threshold T of 2, our algorithm PSA is twice as fast as SA.

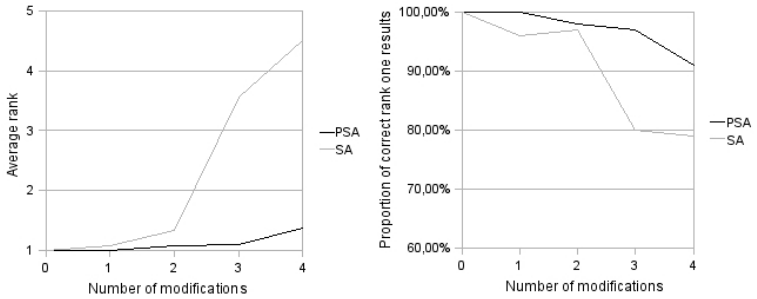


Figure 1. Comparison of SA and PSA on our sets of 1000 random peptides

Number of Possible Masses					
		Threshold T			
		1	2	3	4
Simulated spectra		485	134	39	14
Experimental Maize spectra	*no filtering*	689	312	141	61
	100 most intense peaks	540	180	57	17
	50 most intense peaks	346	79	18	4

Table 1. Evaluation of the number of tested masses on four sets of spectra depending on the threshold T.

3 Conclusion

We have developed PacketSpectralAlignment, a new dynamic programming based algorithm that fully exploits, for the first time, two properties that are inherent to MS/MS spectra. The first one consists in using the *inner symmetry* of spectra and the second one is the grouping of all dependent peaks into *packets*.
Our results are very positive, showing a serious increase in peptides identification in spite of modifications. The sensibility has been significantly increased, while the execution time has been divided by more than two. More tests on experimental data will allow us to evaluate more precisely the benefits provided by our new algorithm. In the future, a better consideration of other points, such as spectra quality, will be added. Moreover, the score will be improved by taking into account other elements such as peaks intensity.

References

[1] P. A. Pevzner, V. Dancík, and C. L. Tang. Mutation-tolerant protein identification by mass spectrometry. *J Comput Biol*, 7(6):777–87, 2000.

[2] P. A. Pevzner, Z. Mulyukov, V. Dancik, and C. L. Tang. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res*, 11(2):290–9, 2001.

Automatic detection of anchor points for multiple alignment

Eduardo Corel[1], Florian Pitschi[2] and Claudine Devauchelle[3]

[1] Institut für Mikrobiologie und Genetik, Goldschmidtstraße 1, 37077 Göttingen, Germany.
ecorel@gwdg.de
[2] Partner Institute for Computational Biology, 320 Yue Yang Rd, 200001 Shanghai, China.
florian.pitschi@googlemail.com
[3] Laboratoire Statistique et Génome CNRS UMR 8071 INRA 1152, Université d'Evry
523 Place des Terrasses, 91034 Evry Cedex, France.
cdevauchelle@genopole.cnrs.fr

1 Introduction

Multiple sequence alignment (MSA) is often a requisite for sequence analysis, but it is a notoriously difficult task. The idea of including local alignment information (pioneered by T-Coffee [1]) is also at work in more recent tools like MUSCLE [2]. A latest trend tends to include structural or homology information retrieved from existing databases (like DbClustal [3]). Yet another way to improve the accuracy of existing MSA is to include user-specified *anchor points*, which are positions that should turn out to be aligned in the output (like in Dialign [4]). We introduce a method to determine automatically a set of anchor points to help multiple alignment software. The anchors are defined combinatorially using a linear complexity algorithm based on the transitive closure of subword composition (N-local decoding [5]), where we adapt locally the value of the length N of the subword. The anchor points produced are naturally multiple, and we define a graph-theoretic algorithm to ensure their consistency with a multiple alignment. Two types of anchor points are introduced into the global software ClustalW2.0 and the improvement is tested on the benchmark BAliBASE3 ([6]).

2 Anchor selection on the partition tree of the local decoding

We consider a collection S of sequences $s = s_1 \dots s_{\ell(s)}$ of lengths $\ell(s)$ over a finite alphabet $\mathcal{A}$. The *site space* $\mathcal{S} = \{(s,p) | s \in S, 1 \leqslant p \leqslant \ell(s)\}$ has a partial ordering $\sigma = (s,p) \leq \sigma' = (s',p')$ if and only if $s = s'$ and $p \leq p'$. We define an *anchor point* as a subset $C \subset \mathcal{S}$ of sites having *at most one occurrence* per sequence. The aim of this whole section is to define a set $\mathcal{C}$ of *disjoint* anchor points. The *succession graph* of a set $\mathcal{C}$ of anchor points is the edge-weighted graph $SG(\mathcal{C}) = (\mathcal{C}, E, w)$ with edges $e = (C, C')$ such that the occurrence of C' follows in some sequence s the occurrence of C, weighed by the number $w(e)$ of sequences where this happens. By convenience we also add an initial vertex v_{start} and a terminal one v_{end}. The set $\mathcal{C}$ is consistent with a global alignment (in the sense of Dialign) if and only if $SG(\mathcal{C})$ is a directed acyclic graph (DAG). A word $w \in \mathcal{A}^N$ *occurs at position i relatively to* $\sigma = (s,p)$ if $s_{[p-i,p-i+N-1]} = w$. Say $\sigma \simeq_N \sigma'$ whenever there is a repeated length N word w at the *same* position relatively to both σ and σ'. The *N-local decoding of $\mathcal{S}$* ([5]) is the partition $\mathcal{E}^N$ of $\mathcal{S}$ induced by the transitive closure of $\simeq_N$. Like any N-mer-based method, no good criterion to tune the parameter N is at hand. However the partitions $\mathcal{E}^N$ are embedded. Letting $\mathcal{E}^0 = \{\mathcal{S}\}$, we can encode the set $V = \bigcup_{i \geq 0} \mathcal{E}^i$ of equivalence classes for different values of N into the *partition tree* $\mathbf{P} = (V, E^{\mathbf{P}})$, where $(u,v) \in E^{\mathbf{P}}$ when $u \in \mathcal{E}^N, v \in \mathcal{E}^{N+1}$ and $v \subsetneq u$. For $C \in \mathcal{E}^N$, let $\kappa(C) = |C|/|\{s \in S | \exists p, (s,p) \in C\}|$. Thus, $\kappa(C) = 1$ whenever C is an anchor point. The **raw** anchors are defined as $\mathcal{C}^{raw}_{\geqslant s_{\min}} = \{C \in V | \kappa(C) = 1$ and $\forall C' \supset C,\ \kappa(C') > 1$ and $|C| \geq$

$s_{\min}\}$. To turn our set of raw anchor points $\mathcal{C} = \mathcal{C}^{raw}_{\geqslant s_{\min}}$ into a consistent one, we remove some sites from the anchor points, so that the succession graph becomes a DAG. We proceed in two steps. First, we delete some edges $e \in E'$ of $SG(\mathcal{C})$ to turn it into a DAG. Finding a set E' of minimal weight is NP-hard, so we remove the lowest weighted edges from $SG(\mathcal{C})$ until all cycles have disappeared. The obtained DAG induces a partial order $\leqslant^*$ on $\mathcal{C} \cup \{v_{start}, v_{end}\}$. For each sequence s, let $\mathcal{C}_s$ be the set of anchor points of $\mathcal{C}$ having a (unique) site (s, j_C) in s. There are two order relations on $V_s = \mathcal{C}_s \cup \{v_{start}, v_{end}\}$, the total order $\leq_s$ induced by the order $\leq$ of $\mathcal{S}$, and the partial order $\leqslant^*_s$ induced by $\leqslant^*$. Let $G_s = (V_s, E_s)$ be the graph of the relation $R = \leq_s \cap \leqslant^*_s$. Choose a path $\gamma_s = (v_{start}, u_1, \ldots, u_n, v_{end})$ of greatest length in G_s from v_{start} to v_{end}. For all anchors $C \in \mathcal{C}_s$ such that $C \notin \gamma_s$, remove the site (s, j_C) from C. The anchor points obtained after performing this procedure for all sequences form a **consistent** anchor set.

3 Results and discussion

On the BAliBASE3 data bank, restricted to datasets having at least 20 sequences to allow for the use of the transitive closure (excluding thus RV10), we have compared the alignments produced by ClustalW2.0 alone and with raw and consistent anchors. The table 1 shows the variation of SP and TC scores computed on the *core regions* by the program `baliscore` from BAliBASE3.

	Δ SP (raw)			Δ SP (consistent)			Δ TC (raw)			Δ TC (consistent)		
$s_{\min}$	all	2–10	11–20	all	2–10	11–20	all	2–10	11–20	all	2–10	11–20
RV20	0.12	-0.14	0.36	0.49	0.51	0.47	-0.02	-1.22	1.07	1.65	2	1.34
RV30	-1.25	-2.7	0.05	1.03	1.43	0.67	3.22	2.5	3.87	4.05	5.53	2.72
RV40	1.29	0.14	2.34	2.2	2.12	2.27	-2.92	-3.88	-2.06	-2.69	-3.14	-2.28
RV50	-0.01	-1.81	1.61	2.05	2.08	2.02	6.03	6.03	5.28	7.75	8.87	6.77

Table 1. Average of the score improvement on ClustalW with both types of anchors.

Introducing the raw anchor points improves slightly both scores for high values of $s_{\min}$, especially TC (with the conspicuous exception of RV40). This means that the total number of correct columns increases noticeably, even though the average number of correctly aligned pairs of residues is almost unchanged, at any rate not diminished. The improvement is clearer (except for TC on RV40 again) if ClustalW receives *consistent* anchors, and the parameter $s_{\min}$ becomes almost irrelevant.

References

[1] Notredame C., Higgins D.G., and Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302: 205–217, 2000.

[2] Edgar R.C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5: 113, 2004.

[3] Thompson J. D., Plewniak F., Thierry J.-C. and Poch O. DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucl. Acids Res.* 28:2919–2926, 2000.

[4] Morgenstern B., Prohaska S., Pöhler D., and Stadler P. F., Multiple sequence alignment with user-defined anchor points. *Algo. Mol. Biol.* 1:6, 2006.

[5] G. Didier, M. Pupin, I. Laprevotte and A. Hénaut. Local decoding of sequences and alignment-free comparison. *J. Comput. Biol.*. Oct;13(8):1465-76, 2006.

[6] Thompson J. D., Plewniak, F., and Poch, O., A comprehensive comparison of multiple sequence alignment programs. *Nucl. Acids Res.* 27:2682–2690, 1999.

Index des auteurs

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z